Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Small area estimation methods under complex sampling designs

Autor:

MARIA GUADARRAMA SANZ

Director:

Isabel Molina Peralta

DEPARTAMENTO DE ESTADÍSTICA

Getafe, Septiembre de 2017

Universidad Carlos III

PH.D. THESIS

Small area estimation methods under complex sampling designs[1]

Author:
María Guadarrama Sanz

Advisor:
Isabel Molina Peralta

DEPARTMENT OF STATISTICS

Getafe, Madrid, October 20, 2017

[1]This work has been supported by the grant MTM2015-69638-R (MINECO/FEDER, UE)

To Mum and Dad

Acknowledgements

A PhD is no easy feat. This dissertation was made possible by the incredible people and organizations with whom I have had the pleasure of working with. Without each of them, my work would not have been possible or as rewarding as it has been.

Firstly, I would like to express my deepest gratitude to my advisor, Prof. Isabel Molina, for her guidance, patience, enthusiasm and immense knowledge which have provided me with a continuous stream of support throughout my PhD and research. Prof. Molina's advice and support have been invaluable. I know for a fact that I could not have had a better advisor and mentor. Also, I am extremely grateful to Prof. Yves Tillé and Prof. J.N.K. Rao for making their endless knowledge available to me.

I also would like to thank my professors from the University of Valladolid who encouraged me to enter the research world, especially Professors Jesús Cavero, Jose Luis Rojo, Isabel Gómez, Jesús Alonso and Pedro Pablo Ortuñez. Similarly, I would like to express my gratitude to my professors from the Department of Statistics of Universidad Carlos III de Madrid, especially Professors Helena Veiga, Juan Miguel Marín, Michael Wiper, Javier Prieto and Regina Kaiser. Thanks to the administratives Susana Linares, Francisco García-Saavedra and Almudena Crespo for making much easier what, without them, would have been administrative nightmares.

I am very grateful to all friends from my village, from Valladolid, as well as those in Getafe who have helped me through the bad times and celebrated with me in the good moments. In particular, I am grateful to Ana and her parents, who make me discover Madrid when I was very young. I thank Ainara and Sandra who, since years ago, have moved together with me and accompanied me during these important stages of my life. Lorena, Yessica and Pablo - friends who, although they may be far away from time to time, are always near in mind. To Pato, because she was my first and best roommate. To Noelia and Robert since they have supported me during these last steps. To colleagues in my department: Daniel and Francisco because four years in the same office deserves recognition; Mario, Diego, Ginette, Andrés, João, Alba, Elisa, Antonio and Hoang, because together we learned that, when life gives you lemons, make lemonade! To the Getagroup girls and boys: Silvina, Luciana, Juliana, Shireen, Marta, Vanesa, Cárita, Facundo, Manu, David, Ab Zereei, Mehdi, Pedro, and Manuel,

because their company has been the very best of my stay in Getafe (as we all know, without Eskinazo, I would never have finished this thesis!).

I would be thoughtless not to express my gratitude to two persons that have been of tremendous help during these years I have been in Getafe: Félix Sanz and Adela Gómez. Without them, Saturday lunches would not have been the same. Moreover, I would like to thank all the people of *Casa Regional de Castilla y León en Getafe*, who brought me a taste of home when I needed it most.

Last but not the least, I would like to thank my parents, whose support has been essential to me throughout my life - especially during my PhD. You taught me to always look forward and never give up. Thanks to the rest of my family: my aunts, uncles, cousins, and grandparents, who have always encouraged me. Grandpa, I miss you!

Abstract

The aim of this thesis is the study of small area estimation methods under outcome-dependent sampling designs, that is, when the selection of the units to the sample depends on their values of the variable of interest. More precisely, we consider two types of informative sampling designs. A first type, in which the inclusion probabilities are strictly positive for all population units and a second type, cut-off sampling, in which a grouping variable related with the variable of interest divides the population in two strata, with one of the strata being deliverately excluded from selection to the sample, that is, where inclusion probabilities are zero. We are specially interested in the estimation of general non-linear parameters, including poverty indicators, in areas or domains of the population with small sample sizes. Due to the small area sample sizes, we will use model-based methods, which borrow strength from all the domains through the assumption of models with common parameters for all the domains.

First, we review the main model-based small area estimation methods for the estimation of general nonlinear parameters, focusing for illustration purposes on particular poverty indicators. We describe direct estimation, which uses data only of the area of study, the empirical best linear unbiased predictor (EBLUP) under the Fay-Herriot at area level model (Fay and Herriot, 1979) and three methods based on unit-level models, namely the method of Elbers et al. (2003) used traditionally by the World Bank, the empirical best/Bayes (EB) method of Molina and Rao (2010) and the hierarchical Bayes proposal of Molina et al. (2014). We put ourselves in the point of view of a practitioner and discuss, as objectively as possible, the benefits and drawbacks of each method, illustrating some of them through simulation studies and also by an application with real data.

In one of the mentioned simulation experiments, we study the performance of the considered estimators under informative sampling. Under informative selection, individuals with certain outcome values appear more often in the sample and, as a consequence, usual inference based on the actual sample without appropriate weighting might be strongly biased. In this dissertation, we propose an extension of the EB method, called pseudo EB (PEB) method, for estimation of general non-linear

parameters in small areas that handles the informative selection by incorporating the sampling weights. We analyze the properties of this method under complex sampling designs, including informative selection. Results confirm that the PEB estimators reduce significantly the bias of unweighted EB estimators under informative sampling, and compare favorably under non-informative selection. We illustrate the procedure through an application to poverty mapping in a Mexican state.

Additionally, we study small area estimation methods under cut-off sampling. This sampling technique consists of excluding a set of units from the selection to the sample due to difficulty in obtaining information from them. In that situation, naïve estimators, obtained by ignoring the cut-off sampling, may be severely design-biased. Calibration estimators using auxiliary information have been proposed to reduce this design-bias. However, the resulting estimators may have large variances when estimating in small domains. Similarly as calibration, model-based small area estimation methods might also help decreasing this bias if the assumed model holds for the whole population. At the same time, these methods provide more efficient estimators than calibration when estimating in small domains. We compare the performance of calibration estimators with the EBLUP or the EB predictors for estimation in small domains under cut-off sampling through simulation studies and a real data application. Our results confirm that the EBLUP under simple random sampling without replacement applied to the non-excluded units helps to reduce the bias due to cut-off sampling. The EBLUP also performs significantly better than naïve direct and calibration estimators in terms of mean squared error. Our results with real data suggest similar conclusions for the EB estimators of nonlinear domain parameters.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Poverty maps are an important source of information on the regional distribution of poverty and are currently used to support regional policy-making and to allocate funds to local jurisdictions. Good examples are the poverty and inequality maps produced by the World Bank for many countries all over the world. In the U.S., the Small Area Income and Poverty Estimates (SAIPE) program of the Census Bureau provides annual estimates of income and poverty statistics for all school districts, counties and states for the administration of federal, state and local programs and the allocation of federal funds to local jurisdictions (https://www.census.gov/did/www/saipe/). In Europe, the joint project "Poverty Mapping in the New Member States of the European Union" between the World Bank and the European Commission was aimed at constructing poverty maps for the new members of the EU. The TIPSE (Territorial Dimension of Poverty and Social Exclusion in Europe) project, commissioned by the European Observation Network for Territorial Development and Cohesion (ESPON) program, aims to support policy by creating a regional database and associated maps of poverty and social exclusion indicators. In Mexico, the National Council for the Assessment of Social Development Policy (CONEVAL in Spanish) is committed by law to produce regular poverty and inequality estimates at the state level by population subgroups and at municipality level. The first objective of the Sustainable Development Goals, substitutes of the former "Millennium Objectives" of the United Nations (UN), is the end of poverty in all its forms everywhere. The UN propose eradicating extreme poverty for all people everywhere by 2030. As an indicator, the UN uses the proportion of population below the international poverty line (currently $1.25 per day) by sex, age, employment status and geographical location, both urban or rural, (https://sustainabledevelopment.un.org/sdg1).

Obtaining accurate poverty maps at high levels of disaggregation is not straightforward because of insufficient sample size of official surveys in some of the target regions. Direct estimates, obtained with the region-specific sample data, are unstable in the sense of having very large sampling errors for regions with small sample size. These unstable poverty estimates might make the seemingly poorer regions in one period appear as the richer in the next period, which can be contradictory. On the other hand, very stable but biased estimates (e.g., too homogeneous across regions) might make identification of the poorer regions difficult.

Sample surveys are the primary data source for official statistics. However, the high cost of interviews leads to an extensive use of the survey data, including estimation for geographical levels (domains) for which they were not initially planned. As already said, when estimating for disaggregated domains, "direct" estimators, based only on the domain-specific sample data, can be highly inefficient because of small area-specific sample sizes. Domains where direct estimators do not have the required precision are called "small areas". Small area estimation methods "borrow strength" by using "indirect" estimators that employ the sample data from other domains, leading to more efficient estimators. Among indirect estimation methods, model-based approaches, which combine survey data with other data sources such as censuses or administrative registers through linking models, are very popular because they can increase considerably the efficiency in very small areas. Pfeffermann et al. (2013) provides a recent review on the topic, and the book by Rao and Molina (2015) contains a comprehensive description of small area estimation methods.

In small area estimation, models are typically classified into area level and unit level models. In area level models, direct area estimators are related to area level auxiliary variables. In unit level models, the values of the study variable for the population units are related to unit-specific covariates. Both unit and area level models have been used extensively to estimate linear parameters such as totals and means. However, many poverty and inequality indicators are complex nonlinear functions of the vector of values of the target variable (e.g. income) in the units of the domain of interest. Specific methods have been developed to address the estimation of general nonlinear parameters in small areas.

Area level models are widely used in official statistics applications. For instance, the U.S. Census Bureau uses the Fay-Herriot (FH) area level model (Fay and Herriot, 1979) to produce model-based county estimates of the number of school-age children under poverty. Molina and Morales (2015) also used the FH model to estimate poverty incidences and gaps in Spanish provinces. The first unit level model designed to estimate general non-linear parameters such as poverty indicators in small areas was

proposed by Elbers et al. (2003), and will hereafter be called ELL method. This method has been extensively used by the World Bank to obtain disaggregated poverty and inequality maps of many countries. A more recent approach is the empirical best/Bayes (EB) method proposed by Molina and Rao (2010). This method also uses a unit level model and delivers the "best" (or optimal) estimator in the sense of minimizing the mean squared error (MSE) under the assumed model. Molina et al. (2014) have proposed a hierarchical Bayes (HB) analogue to the EB method to estimate general non-linear parameters in small areas.

In many applications, individuals with certain outcome values are more likely selected for the sample. For example, in forestry, the largest trees may be more likely to be selected; in case-control studies, cases are typically selected with larger probability than controls. In such situations, we say that the selection mechanism (or sampling design) is informative. The result of an informative selection is a sample that "without appropriate weighting" is not representative of the target population. In this situation, estimators which do not take into account the design weights may display large bias. Thus, a weighting procedure is needed to downweight the outcomes of individuals that appear more often in the sample. This is the idea of design-based estimation, where the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), also called "expansion estimator" and the Hajék estimator (Hájek, 1971), which is simply the weighted sample mean (WSM), are the basic estimators of a population mean. These design-based estimators do not require model assumptions and are consistent under general sampling mechanisms as the sample size becomes large. Model-based methods that are obtained ignoring the sampling weights, such as the EBLUP and the EB, may display a large bias under informative selection.

An extreme type of informative selection is cut-off sampling, where a set of population units are not accounted for the selection of the sample. This type of sampling is employed in many business surveys, where small firms are excluded from the sample because of the high cost of obtaining and maintaining a frame covering the whole population of firms. According to Särndal et al. (1992), pp. 531-533, this sampling technique is used when the distribution of the study variable is highly skewed and there is not a reliable frame covering the small elements. Benedetti et al. (2010) enumerate the practical advantages of cut-off sampling in terms of the survey reduction cost. The monthly survey of manufacturing performed by Statistics Canada is an example of cut-off sampling (Benedetti et al., 2010). In Spain, the monthly survey of industrial production index (IPI) performed by the Spanish National Statistical Institute (in Spanish, INE) collects data from those firms which produce a significant volume of products according to the annual industrial survey of products (in Spanish, EIAP), see

http://www.ine.es/daco/daco43/metoipi10.pdf. Related surveys, e.g. the Index of Industrial Prices (IIP) and the Index of Business Turnover (IBT) also use this sampling technique.

Cut-off sampling leads to biased estimates since the inclusion probabilities for excluded units are zero. Consecuently, the sampling weights for those units do not exist, see e.g. Särndal et al. (1992) and Haziza et al. (2010) among others. Haziza et al. (2010) propose to use auxiliary information either at the design or at the estimation stage in order to reduce the bias when estimating population totals; more specifically, they propose to use balanced sampling or calibration, or both. Here, we restrict ourselves to the estimation stage. At this stage, Haziza et al. (2010) propose to use auxiliary information through calibration. Calibration estimators, or generalized regression (GREG) estimators, provide good results in terms of reduction of bias and efficiency when estimating at national level or at low levels of dissagregation where the sample size is large enough. However, for domains with small sample size, calibration estimators fail in terms of efficiency, displaying large variances. Model-based small area estimation methods, which use auxiliary information as well, reduce the bias due to cut-off sampling similarly as calibration. Moreover, because of the increase of the "effective" sample size, small area estimators help to reduce the variance in domains with small sample sizes. Furthermore, these methods allow the estimation of more complex parameters.

1.2 Organization and outline of the thesis

The Ph. Dissertation is organized as follows. In Chapter 2, we review the main methods for estimation of general parameters in small areas. We consider estimators based on area level models such as Fay Herriot model (Fay and Herriot, 1979) and estimators based on unit level models, such as ELL method of Elbers et al. (2003), the best/Bayes estimator (EB) of Molina and Rao (2010), and the hierarchical Bayes estimator (HB) of Molina et al. (2014). We comment on the advantages and disadvantages of these methods from a practical point of view, illustrating them through simulation studies under different scenarios. First, we consider the case of the sample drawn by simple random sampling without replacement (srswor), then the case of informative sampling with different degrees of informativeness and, finally, we illustrate the case when outliers with different levels and intensities are present in the data. We construct poverty maps of Spain by provinces.

Chapter 3 is concerned with our proposed approach to reduce the bias due to informative selection of the sample. Since the bias showed by the unweighted

estimators such as EB is due to ignoring the sampling mechanism, we propose a weighted version of EB method, called pseudo EB (PEB). This procedure is based on the same conditioning idea of EB, but in this case we condition on the domain weighted sample mean (WSM). Simulation studies indicate a large reduction of the relative bias and good performance in terms of efficiency of pseudo EB estimators compared to EB. We prove the approximately design consistency of the PEB estimators of poverty incidence and gap. We illustrate the use of the proposed method estimating the poverty incidence and gap for municipalities of the State of Mexico.

Chapter 4 deals with cut-off sampling. We compare the already existing proposals with small area estimation methods. The main characteristic of cut-off sampling is that part of the population is excluded from possible sample selection. Naïve estimators obtained by ignoring this fact lead to biased estimators. The solutions proposed so far, which use auxiliary information through calibration, may show large variances when estimating in small domains. We propose small area estimation methods to gain efficiency. Precisely, we consider the EBLUP in the case of linear parameters and the EB estimator for more general parameters. We compare the performance of calibration and the mentioned small area estimation methods through simulation studies. Results show that EBLUP and calibration estimates significantly reduce the bias of direct estimators due to cut-off sampling but, in terms of efficiency, calibration estimators exhibit large mean squared error in areas with small sample sizes. We use calibration and EB estimators to estimate the total sales of a specific tobacco product in Spanish provinces. Finally, Chapter 5 draws some conclusions from the thesis and proposes future research lines.

To make it easy for the reader, we have tried to make each chapter completely self-contained even if falling in the danger of becoming repetitive. Concerning notation, Chapters 2-3 try to follow the same notation, but some notation is changed in Chapter 4, trying to respect the conventional notation of related books and manuscripts. Even if some of the symbols have different meaning as in previous chapters, every symbol in this chapter is newly defined and we believe that all concepts are thus kept clear.

Chapter 2

A comparison of small area estimation methods for poverty mapping

In the literature, we can find many different indicators describing wellbeing of people. The FGT class of poverty indicators introduced by Foster et al. (1984) includes basic indicators such as the at-risk-of-poverty rate, also called poverty incidence and defined as the proportion of individuals with welfare below the poverty line and the poverty gap defined as the mean relative distance to the poverty line. Poverty indicators that do not require definition of a poverty line are the Sen Index, Fuzzy monetary and Fuzzy supplementary poverty indicators (Betti et al., 2006). Inequality measures include Quintile Share Ratio, Gini index, Theil index, the Generalized entropy class (Theil index belongs to this class) and Atkinson's inequality measures. For a description of these measures see e.g. Neri et al. (2005). As part of the Lisbon strategy of 2000, which envisioned the coordination of European social policies at country level based on a set of common goals, the European Council of Dec. 2001 established a set of common European statistical indicators on poverty and social exclusion called Laeken indicators (Stubbs et al., 2008). This set contains several of the indicators mentioned above like the at-risk-of-poverty rate, the Quintile Share Ratio and Gini index.

In this chapter, we review the main methods for the estimation of general non-linear small domain parameters, focusing for illustrative purposes on the FGT family of poverty indicators, which is introduced in Section 2.1. Specifically, in Section 2.2, we describe direct estimation, the EBLUP based on the Fay and Herriot (1979) area level model used by the U.S. Census Bureau, the method of Elbers et al. (2003) used by the World Bank, the more recent empirical Best/Bayes (EB) method of Molina and Rao

(2010) together with its variation called Census EB, and the hierarchical Bayes (HB) method of Molina et al. (2014). We discuss advantages and disadvantages of each procedure from a practical point of view. In Section 2.3, we illustrate their performance in simulations under several scenarios, including the cases of informative sampling or the presence of outliers. Finally, Section 2.4 applies several of the introduced procedures to poverty mapping in Spanish provinces by gender.

2.1 Poverty indicators

Although the methods reviewed in this chapter can be applied to many different poverty and inequality indicators, for simplicity of exposition and illustrative purposes, we will focus on the FGT family of poverty indicators. Consider a population U of size N that is partitioned into m domains or areas $U_1, \ldots, U_m$, of sizes $N_1, \ldots, N_m$. Let E_{ij} be a measure of welfare for individual j ($j = 1, \ldots, N_i$) in domain i ($i = 1, \ldots, m$). Let z be the poverty line, that is, the value such that when $E_{ij} < z$, individual j from domain i is regarded as "at risk of poverty". Then, the FGT family of poverty indicators for domain i is given by

$$F_{\alpha i} = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z - E_{ij}}{z} \right)^{\alpha} I(E_{ij} < z), \quad \alpha \geq 0,\ i = 1, \ldots, m, \tag{2.1}$$

where $I(E_{ij} < z) = 1$ if $E_{ij} < z$ and $I(E_{ij} < z) = 0$ otherwise. For $\alpha = 0$, we obtain the proportion of individuals "at risk of poverty", that is, the poverty incidence or at-risk-of-poverty rate. For $\alpha = 1$, we get the average of the relative distances to non being "at risk of poverty", called poverty gap. The poverty incidence measures the frequency of poverty, whereas the poverty gap measures the intensity of poverty.

We remark that the unit level methods introduced in this chapter can be applied to estimate any desired population characteristic that is obtained as a real measurable function of the values of a continuous variable in the units of the population, as long as this variable follows the considered model in each method.

2.2 Estimators

Estimation of population characteristics is typically based on a sample s drawn from the population U. We denote by $s_i = s \cap U_i$ the subsample from domain i of size $n_i < N_i$ and by $r_i = U_i - s_i$ the complement of s_i, of size $N_i - n_i$. The overall sample size is $n = \sum_{i=1}^{m} n_i, i = 1, \ldots, m$. The following subsections describe common estimators of poverty indicators obtained from the sample data.

2.2.1 Direct estimators

When estimating in a given domain or area i, a direct estimator uses only the n_i observations from that domain, provided that this domain has been sampled (i.e., $n_i > 0$). The FGT poverty indicator (2.1) of order α for domain i can be expressed as a mean as follows

$$F_{\alpha i} = N_i^{-1} \sum_{j=1}^{N_i} F_{\alpha ij}, \quad F_{\alpha ij} = \left(\frac{z - E_{ij}}{z} \right)^{\alpha} I(E_{ij} < z), \quad j = 1, \ldots, N_i.$$

Since $F_{\alpha i}$ is now a mean of the domain elements $F_{\alpha ij}$, we can easily obtain the Horwitz-Thompson (HT) direct estimator of $F_{\alpha i}$ as

$$\hat{F}_{\alpha i} = N_i^{-1} \sum_{j \in s_i} w_{ij} F_{\alpha ij}, \tag{2.2}$$

where $w_{ij} = \pi_{ij}^{-1}$ is the sampling weight of unit j from domain i and π_{ij} is the inclusion probability of unit j in the subsample s_i. The idea of the HT estimator (2.2) is to downweight observations with larger probability of appearing in the sample and upweight those with smaller probability. Under the assumption that the second-order inclusion probabilities within domain i, π_{jl}^i, satisfy $\pi_{jl}^i = \pi_j^i \pi_l^i$, $j \neq l$, which holds exactly under Poisson sampling within domain i, a design-unbiased estimator of the design variance of $\hat{F}_{\alpha i}$, $\hat{V}_\pi(\hat{F}_{\alpha i})$, is given by,

$$\hat{V}_\pi(\hat{F}_{\alpha i}) = N_i^{-2} \sum_{j \in s_i} w_{ij}(w_{ij} - 1) F_{\alpha ij}^2. \tag{2.3}$$

Below, we list the advantages and disadvantages of direct estimators, such as the HT estimator (2.2), for small area estimation.

Advantages:

- They are (at least approximately) design-unbiased and design-consistent as $n_i \to \infty$. Thus, they perform well under complex sampling designs, including informative sampling, as long as they are calculated using the correct inclusion probabilities.
- They do not require model assumptions; that is, direct estimators are completely nonparametric.

Disadvantages:

- They are very inefficient for domains with very small sample size n_i.

- They cannot be calculated for nonsampled domains (i.e., with $n_i = 0$).

2.2.2 Fay-Herriot model

FH area level model was introduced by Fay and Herriot (1979) and is currently used by the U.S. Census Bureau within the Small Area Income and Poverty Estimates (SAIPE) program (https://www.census.gov/did/www/saipe/). This model links the parameters of interest for all the domains, $F_{\alpha i}$, $i = 1, \ldots, m$, through a linear model as

$$F_{\alpha i} = \mathbf{x}_i'\boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m, \tag{2.4}$$

where $\mathbf{x}_i$ is a p-vector of area level covariates, $\boldsymbol{\beta}$ is the regression parameter common for all domains, and v_i is the domain-specific regression error measuring the unexplained domain heterogeneity and also called random effect for domain i. We assume that domain random effects v_i are independent and identically distributed (iid), with unknown variance σ_v^2, that is, $v_i \overset{iid}{\sim} (0, \sigma_v^2)$. Note that true values $F_{\alpha i}$ are not observable and therefore model (2.4) cannot be directly fitted. However, we can make use of a direct estimator $\hat{F}_{\alpha i}$ of $F_{\alpha i}$. FH model assumes that $\hat{F}_{\alpha i}$ is design-unbiased, with

$$\hat{F}_{\alpha i} = F_{\alpha i} + e_i, \quad i = 1, \ldots, m, \tag{2.5}$$

where e_i is the sampling error for domain i. We assume that sampling errors e_i are independent of domain effects v_i and satisfy $e_i \overset{ind}{\sim} (0, \psi_i)$, where the sampling variances ψ_i, $i = 1, \ldots, m$, are assumed to be known. Combining (2.4) and (2.5), we obtain a linear mixed model

$$\hat{F}_{\alpha i} = \mathbf{x}_i'\boldsymbol{\beta} + v_i + e_i, \quad i = 1, \ldots, m. \tag{2.6}$$

The best linear unbiased predictor (BLUP) of $F_{\alpha i} = \mathbf{x}_i'\boldsymbol{\beta} + v_i$ under model (2.6) is simply obtained by fitting model (2.6) and replacing the estimate of $\boldsymbol{\beta}$ and the predictor of v_i in (2.6), that is, the BLUP of $F_{\alpha i}$ is given by

$$\tilde{F}_{\alpha i}^{FH} = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{v}_i, \tag{2.7}$$

where $\tilde{v}_i = \gamma_i(\hat{F}_{\alpha j} - \mathbf{x}_i'\tilde{\boldsymbol{\beta}})$ is the BLUP of v_i, with $\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$ and where $\tilde{\boldsymbol{\beta}}$ is the weighted least squares estimator of $\boldsymbol{\beta}$, given by

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \gamma_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{m} \gamma_i \mathbf{x}_i \hat{F}_{\alpha i}.$$

see Appendix A.1. In practice, the variance σ_v^2 of the domain effects v_i is unknown and needs to be estimated. Common estimation methods are maximum likelihood (ML) and restricted maximum likelihood (REML). REML corrects for the degrees of freedom due to estimating $\boldsymbol{\beta}$ and leads to a less biased estimator of σ_v^2 for finite sample size n. Let $\hat{\sigma}_v^2$ be the resulting estimator. Replacing $\hat{\sigma}_v^2$ for σ_v^2 in (2.7), we obtain the empirical BLUP (EBLUP) of $F_{\alpha i}$ based on FH model (2.6), denoted here as $\hat{F}_{\alpha i}^{FH}$ and called simply FH estimator. Under normality of v_i and e_i, an approximation up to $o(m^{-1})$ terms for the MSE of the FH estimator was obtained by Prasad and Rao (1990) and is given by

$$\text{MSE}(\hat{F}_{\alpha i}^{FH}) = g_{i1}(\sigma_v^2) + g_{i2}(\sigma_v^2) + g_{i3}(\sigma_v^2),$$

where

$$\begin{aligned} g_{1i}(\sigma_v^2) &= \gamma_i \psi_i, \\ g_{2i}(\sigma_v^2) &= (1-\gamma_i)^2 \mathbf{x}_i' \left(\sum_{i=1}^{m} \gamma_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \mathbf{x}_i, \\ g_{3i}(\sigma_v^2) &= (1-\gamma_i)^2 (\sigma_v^2 + \psi_i^2)^{-1} \bar{V}_\pi(\hat{\sigma}_v^2). \end{aligned}$$

Here, $\bar{V}_\pi(\hat{\sigma}_v^2)$ is the asymptotic variance of the estimator $\hat{\sigma}_v^2$ of σ_v^2 under the assumed model in this case (2.6).

Good and bad properties of FH estimator (2.7) are listed below, including particular properties for poverty mapping.

Advantages:

- The BLUP under FH model can be expressed as a weighted combination of the direct and the regression-synthetic estimators, that is,

$$\tilde{F}_{\alpha i}^{FH} = \gamma_i \hat{F}_{\alpha i} + (1-\gamma_i) \mathbf{x}_i' \tilde{\boldsymbol{\beta}}, \quad i = 1, \ldots, m. \tag{2.8}$$

  with weight $\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$. Then, for a domain i in which the direct estimator $\hat{F}_{\alpha i}$ is inefficient, that is, with a large sampling variance ψ_i compared to the unexplained between-domain variability σ_v^2, γ_i becomes small and $\tilde{F}_{\alpha i}^{FH}$ borrows more strength from the other domains through the regression synthetic estimator $\mathbf{x}_i' \tilde{\boldsymbol{\beta}}$. On the other hand, for a domain i in which the direct estimator $\hat{F}_{\alpha i}$ is efficient, that is, with small sampling variance ψ_i compared to the unexplained between-domain variability σ_v^2, γ_i is large and $\tilde{F}_{\alpha i}^{FH}$ attaches more weight to the direct estimator. Thus, FH estimator automatically borrows strength for the domains where it is actually needed.

- If $\gamma_i > 0$ for domain i, it makes use of the sampling weights w_{ij} through the direct estimator $\hat{F}_{\alpha i}$. Thus, it is design-consistent as $n_i \to \infty$. As a consequence, it is less affected by informative sampling provided that the direct estimator is calculated using the correct inclusion probabilities.
- Due to the aggregation of data, it is not very much affected by isolated unit level outliers.
- It requires only domain level auxiliary information and therefore avoids the confidentiality issues associated with micro-data.

Disadvantages:

- The sampling variances ψ_i are assumed to be known, but in practice they are estimated. It is not easy to incorporate the uncertainty due to estimation of the sampling variances in the MSE.
- The number of observations used to fit the FH model is the number of domains m, which is typically much smaller than the number of observations used to fit unit level models, n. Thus, model parameters are estimated with less efficiency and, therefore, the efficiency gains with respect to direct estimators are expected to be smaller than under unit level models.
- It requires normality of v_i and e_i for MSE estimation. This might not hold for very complex poverty indicators.
- If we want to estimate several indicators depending on a common continuous variable, it requires separate modeling and searching for good covariates for each indicator.
- Once the model is fitted at the domain level, small area estimates $\hat{F}_{\alpha i}^{FH}$ cannot be further disaggregated for subdomains or subareas within the domains unless a new good model is found at that subdomain level.

2.2.3 ELL method

ELL method assumes a unit level linear mixed model for a log-transformation of the variable measuring welfare of individuals, with random effects for the sampling clusters or primary sampling units. For comparability with the rest of the methods presented here, in the following, we assume that the sampling clusters are the domains. In this case, the model becomes the nested error model of Battese et al. (1988) for the log-transformation of the welfare variables, that is, $Y_{ij} = \log(E_{ij})$ is assumed to be linearly related with a p-vector of auxiliary variables $\mathbf{x}_{ij}$, which may include

unit-specific and domain-specific covariates. The model also includes random area effects v_i as follows

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1, \ldots, N_i,\ i = 1, \ldots, m. \tag{2.9}$$

Here, $\boldsymbol{\beta}$ is a p-vector of regression coefficients, $v_i \overset{iid}{\sim} (0, \sigma_v^2)$, $e_{ij} \overset{ind}{\sim} (0, \sigma_e^2 k_{ij}^2)$ where v_i and e_{ij} are independent, and k_{ij} are known constants that may account for heteroscedasticity.

ELL estimator of $F_{\alpha i}$ is given by the marginal expectation under model (2.9) $\hat{F}_{\alpha i}^{ELL} = E_m[F_{\alpha i}]$. This estimator, together with its MSE under model (2.9), are approximated by a bootstrap method. In this bootstrap procedure, random effects v_i^* and model errors e_{ij}^* are generated from residuals obtained by fitting model (2.9) to survey data. Then, a bootstrap census of Y-values is generated as

$$Y_{ij}^* = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + v_i^* + e_{ij}^*, \quad j = 1, \ldots, N_i,\ i = 1, \ldots, m,$$

where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$. The generation is repeated for $a = 1, \ldots, A$, obtaining A censuses. Then, for each bootstrap census a, the FGT poverty indicator for domain i is calculated as

$$F_{\alpha i}^{*(a)} = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z - \exp(Y_{ij}^{*(a)})}{z} \right)^{\alpha} I(\exp(Y_{ij}^{*(a)}) < z).$$

The ELL estimator of $F_{\alpha i}$ is then approximated by averaging over the A generated censuses, that is,

$$\hat{F}_{\alpha i}^{ELL} = \frac{1}{A} \sum_{a=1}^{A} F_{\alpha i}^{*(a)}.$$

The MSE of $\hat{F}_{\alpha i}^{ELL}$ under model (2.9), $\text{MSE}_m(\hat{F}_{\alpha i}^{ELL})$, is then estimated in Elbers et al. (2003) as follows

$$mse_m(\hat{F}_{\alpha i}^{ELL}) = \frac{1}{A} \sum_{a=1}^{A} \left(F_{\alpha i}^{*(a)} - \hat{F}_{\alpha i}^{ELL} \right)^2. \tag{2.10}$$

Advantages and disadvantages of ELL method are listed below:

Advantages:

- It is based on unit level data, which are richer than area level data and typically uses much larger sample size (n compared to m) to fit the model.

- ELL method can be applied to estimate general indicators defined as function of the model response variables Y_{ij}.
- They are model-unbiased if the model parameters are known.
- Once the model is fitted, estimates can be obtained at whatever subdomain level.

Disadvantages:

- In terms of model MSE, ELL estimates perform poorly and can even perform worse than direct estimators when unexplained between-domain variation is significant, see Molina and Rao (2010). In fact, for the estimation of domain means, ELL estimates are basically equal to regression-synthetic estimators, which assume the regression model without further between-domain variation.
- They are based on a model assumption. Hence, model checking is crucial.
- They are not design-unbiased and can be seriously biased under informative sampling.
- They can be seriously affected by unit level outliers.
- If cluster effects are included in the model instead of area effects, but area effects are significant, ELL estimates of the model MSE can seriously underestimate the true MSE. Even if area effects are included in the model, ELL estimates of MSE do not track correctly the true model MSE for each domain.

2.2.4 Empirical Best/Bayes (EB) method

The EB method of Molina and Rao (2010) assumes that the population variables Y_{ij} follow the nested error model (2.9) with normality of random effects v_i and errors e_{ij}. Under that model, the domain vectors $\mathbf{y}_i = (Y_{i1}, \ldots, Y_{iN_i})'$ are independent for $i = 1, \ldots, m$ and satisfy $\mathbf{y}_i \overset{ind}{\sim} N(\boldsymbol{\mu}_i, \mathbf{V}_i)$, where $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i})'$ and $\mathbf{V}_i = \sigma_v^2 \mathbf{1}_{N_i}\mathbf{1}_{N_i}' + \sigma_e^2 \mathbf{A}_i$, for $\mathbf{A}_i = \text{diag}(k_{ij}^2; j = 1, \ldots, N_i)$. For a domain parameter $H_i = h(\mathbf{y}_i)$, the estimator that minimizes the MSE, called the best estimator, is given by

$$\tilde{H}_i^B = E_{\mathbf{y}_{ir}}[h(\mathbf{y}_i)|\mathbf{y}_{is}; \boldsymbol{\theta}] = \int h(\mathbf{y}_i) f(\mathbf{y}_{ir}|\mathbf{y}_{is}; \boldsymbol{\theta}) d\mathbf{y}_{ir}, \tag{2.11}$$

where $f(\mathbf{y}_{ir}|\mathbf{y}_{is}; \boldsymbol{\theta})$ is the conditional probability density function (pdf) of the vector $\mathbf{y}_{ir}$ of out-of-sample values, Y_{ij}, $j \in r_i$, from domain i given the vector $\mathbf{y}_{is}$ of sample values, Y_{ij}, $j \in r_i$ from domain i, and $\boldsymbol{\theta}$ is the vector of model parameters. Now replacing $\boldsymbol{\theta}$ in (2.11) by a consistent estimator $\hat{\boldsymbol{\theta}}$, we get the empirical best estimator, $\hat{H}_i^{EB}$.

Under the nested error model (2.9), the distribution of $\mathbf{y}_{ir}|\mathbf{y}_{is}$ is easy to derive. First, we decompose $\mathbf{X}_i$ and $\mathbf{V}_i$ into sample and out-of-sample elements similarly as we do with $\mathbf{y}_i$, that is,

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{is} \\ \mathbf{y}_{ir} \end{pmatrix}, \quad \mathbf{X}_i = \begin{pmatrix} \mathbf{X}_{is} \\ \mathbf{X}_{ir} \end{pmatrix}, \quad \mathbf{V}_i = \begin{pmatrix} \mathbf{V}_{is} & \mathbf{V}_{isr} \\ \mathbf{V}_{irs} & \mathbf{V}_{ir} \end{pmatrix}.$$

By the normality assumption, we have that $\mathbf{y}_{ir}|\mathbf{y}_{is} \overset{ind}{\sim} N(\boldsymbol{\mu}_{ir|s}, \mathbf{V}_{ir|s})$, where the conditional mean vector and covariance matrix are given by

$$\boldsymbol{\mu}_{ir|s} = \mathbf{X}_{ir}\boldsymbol{\beta} + \gamma_{ic}(\bar{y}_{ic} - \bar{\mathbf{x}}'_{ic}\boldsymbol{\beta})\mathbf{1}_{N_i-n_i}, \tag{2.12}$$

$$\mathbf{V}_{ir|s} = \sigma_v^2(1-\gamma_i)\mathbf{1}_{N_i-n_i}\mathbf{1}'_{N_i-n_i} + \sigma_e^2 \text{diag}_{j\in r_i}(k_{ij}^2). \tag{2.13}$$

Here, $\gamma_{ic} = \sigma_v^2(\sigma_v^2 + \sigma_e^2/c_{i\cdot})^{-1}$, for $c_{i\cdot} = \sum_{j\in s_i} c_{ij}$ with $c_{ij} = k_{ij}^{-2}$, and $\bar{y}_{ic}$ and $\bar{\mathbf{x}}_{ic}$ are weighted sample means obtained as

$$\bar{y}_{ic} = \frac{1}{c_{i\cdot}} \sum_{j\in s_i} c_{ij} Y_{ij}, \quad \bar{\mathbf{x}}_{ic} = \frac{1}{c_{i\cdot}} \sum_{j\in s_i} c_{ij}\mathbf{x}_{ij}. \tag{2.14}$$

In practice, for complex non-linear parameters $H_i = h(\mathbf{y}_i)$, the expectation given in (2.11) cannot be calculated analytically and it is approximated by Monte Carlo. This requires to simulate multivariate Normal vectors $\mathbf{y}_{ir}^{(a)}$ of sizes $N_i - n_i$, $i = 1, \ldots, m$, from the (estimated) conditional distribution of $\mathbf{y}_{ir}|\mathbf{y}_{is}$ and then to replicate for $a = 1, \ldots, A$, which may be computationally unfeasible. This can be avoided by noting that the conditional covariance matrix $\mathbf{V}_{ir|s}$, given by (2.13), corresponds to the covariance matrix of a random vector $\mathbf{y}_{ir}^{(a)}$ generated from the model

$$\mathbf{y}_{ir}^{(a)} = \boldsymbol{\mu}_{ir|s} + v_i^{(a)}\mathbf{1}_{N_i-n_i} + \boldsymbol{\epsilon}_{ir}^{(a)}, \tag{2.15}$$

where $v_i^{(a)}$ and $\boldsymbol{\epsilon}_{ir}^{(a)}$ are independent and satisfy

$$v_i^{(a)} \sim N(0, \sigma_v^2(1-\gamma_{ic})) \quad \text{and} \quad \boldsymbol{\epsilon}_{ir}^{(a)} \sim N(\mathbf{0}_{N_i-n_i}, \sigma_e^2 \text{diag}_{j\in r_i}(k_{ij}^2));$$

see Molina and Rao (2010). Using model (2.15), instead of generating a multivariate normal vector $\mathbf{y}_{ir}^{(a)}$ of size $N_i - n_i$, we just need to generate $1 + N_i - n_i$ independent univariate normal variables $v_i^{(a)} \overset{ind}{\sim} N(0, \sigma_v^2(1-\gamma_i))$ and $\epsilon_{ij}^{(a)} \overset{ind}{\sim} N(0, \sigma_e^2 k_{ij}^2)$, for $j \in r_i$. Then, we obtain the corresponding out-of-sample values $Y_{ij}^{(a)}$, $j \in r_i$, from (2.15) using as means, the corresponding elements $\mu_{ij|s}$ of $\boldsymbol{\mu}_{ir|s}$ given by (2.12). Using the vector $\mathbf{y}_{ir}^{(a)}$ generated from (2.15), we construct the census vector $\mathbf{y}_i^{(a)} = (\mathbf{y}'_{is}, (\mathbf{y}_{ir}^{(a)})')'$ and calculate

the parameter of interest $H_i^{(a)} = h(\mathbf{y}_i^{(a)})$. For a non-sampled domain i (i.e., with $n_i = 0$), we generate $\mathbf{y}_{ir}^{(a)}$ from (2.15) with $\gamma_{ic} = 0$ and in this case $\mathbf{y}_i^{(a)} = \mathbf{y}_{ir}^{(a)}$. The Monte Carlo approximation to the EB estimator (2.11) of $H_i = h(\mathbf{y}_i)$ is then given by

$$\hat{H}_i^{EB} \approx \frac{1}{A}\sum_{a=1}^{A} h(\mathbf{y}_i^{(a)}). \tag{2.16}$$

In particular, to estimate the FGT poverty indicator given in (2.1), Molina and Rao (2010) assumed that $Y_{ij} = T(E_{ij})$ follow the nested error model (2.9), where Y_{ij} is the transformed welfare and $T(\cdot)$ is a one-to-one transformation. In terms of the vector of transformed variables $\mathbf{y}_i = (Y_{i1}, \ldots, Y_{iN_i})'$, the FGT poverty indicator can be expressed as

$$F_{\alpha i} = \frac{1}{N_i}\sum_{j=1}^{N_i}\left(\frac{z - T^{-1}(Y_{ij})}{z}\right)^{\alpha} I(T^{-1}(Y_{ij}) < z) = h_\alpha(\mathbf{y}_i), \tag{2.17}$$

and the above EB method can be applied to the domain parameter $H_i = h_\alpha(\mathbf{y}_i)$.

In the case of complex parameters such as the FGT poverty indicators, analytic approximations for the MSE are hard to derive. Molina and Rao (2010) obtained a parametric boostrap MSE estimator following the bootstrap method for finite populations of González-Manteiga et al. (2008), see Molina and Rao (2010) for further details.

Note that both ELL and EB methods require a survey data file containing the observations from the target variable and the auxiliary variables, that is, $\{(Y_{ij}, \mathbf{x}_{ij}); j \in s_i,\ i = 1, \ldots, m\}$, and a census containing the values of the same auxiliary variables for all the units in the population, that is, $\{\mathbf{x}_{ij}; j = 1, \ldots, N_i,\ i = 1, \ldots, m\}$. EB method requires additionally to identify the set of out-of-sample units r (or equivalently the sample units s) in the census U. Linking the survey and the census files is not always possible in practice. However, typically the domain sample size n_i is really small compared to the population size N_i. Then, we can use the Census EB estimator described in Guadarrama et al. (2016), and obtained by generating in each Monte Carlo replicate the full census vector $\mathbf{y}_i$ rather than only the vector of out-of-sample observations $\mathbf{y}_{ir}$. For this, we apply the Monte Carlo approximation (2.16) by generating $\mathbf{y}_i^{(a)} = \boldsymbol{\mu}_{i|s} + v_i^{(a)}\mathbf{1}_{N_i - n_i} + \boldsymbol{\epsilon}_i^{(a)}$, where $\boldsymbol{\mu}_{i|s} = \mathbf{X}_i\boldsymbol{\beta} + \gamma_{ic}(\bar{y}_{ic} - \bar{\mathbf{x}}_{ic}'\boldsymbol{\beta})\mathbf{1}_{N_i}$ and $\boldsymbol{\epsilon}_i^{(a)} \sim N(\mathbf{0}_{N_i}, \sigma_e^2 \text{diag}_{j=1,\ldots,N_i}(k_{ij}^2))$. If the sampling fraction n_i/N_i is negligible, the Census EB estimator of $H_i = F_{\alpha i}$ is practically the same as the original EB estimator.

Good properties and drawbacks of the EB method are listed below.

Advantages:

- It is based on unit level data, which are richer than the area level data and uses

much larger sample size to fit the model.

- EB method can be applied to estimate general indicators defined as functions of the response variables Y_{ij}.
- Best estimators are exactly model-unbiased.
- EB estimators are optimal in terms of minimizing the model MSE for known values of model parameters.
- EB estimators perform significantly better than ELL estimates when unexplained between-domain variation is significant. For out-of-sample domains (with $n_i = 0$), EB and ELL small area estimates are nearly the same. They are nearly the same for all domains if there is no unexplained between-domain variation ($\sigma_v^2 = 0$).
- Once the model is fitted, estimates can be obtained at whatever subdomain level.

Disadvantages:

- They are based on a model assumption. Hence, model checking is crucial.
- They are not approximately design-unbiased and can be seriously biased under informative sampling.
- They can be severely affected by unit level outliers.
- Parametric bootstrap estimates of the MSE of EB estimators are computationally intensive.

2.2.5 Hierarchical Bayes (HB) method

Computation of EB (and Census EB) estimates supplemented with their MSE estimates is very intensive and might be unfeasible for very large populations or for very complex indicators. Note that to approximate the EB estimate by Monte Carlo, we need to construct a large number A of censuses $\mathbf{y}^{(a)}$, where each one might be of huge size. Moreover, to obtain the parametric bootstrap MSE estimator, the Monte Carlo approximation needs to be repeated for each bootstrap replicate. Seeking for a computationally more efficient approach, Molina et al. (2014) developed the alternative HB method for estimation of complex non-linear parameters. This approach does not require the use of bootstrap for MSE estimation because it provides samples from the posterior distribution, from which posterior variances play the role of MSEs, and any other useful posterior summary can be easily obtained.

The HB method is based on reparameterizing the nested error model (2.9) in terms of the intraclass correlation coefficient $\rho = \sigma_v^2/(\sigma_v^2 + \sigma_e^2)$ and considering only

non-informative priors for the model parameters $(\boldsymbol{\beta}, \rho, \sigma_e^2)$. Concretely, the HB model is defined as

$$\begin{aligned}
&\text{(i)} \quad Y_{ij}|v_i, \boldsymbol{\beta}, \sigma_e^2, \rho \overset{ind}{\sim} N(\mathbf{x}_{ij}'\boldsymbol{\beta} + v_i, \sigma_e^2 k_{ij}^2), \quad j = 1, \ldots, N_i, \\
&\text{(ii)} \quad v_i|\rho, \sigma_e^2 \overset{iid}{\sim} N\left(0, \frac{\rho}{1-\rho}\sigma_e^2\right), \quad i = 1, \ldots, m, \\
&\text{(iii)} \quad \pi(\boldsymbol{\beta}, \rho, \sigma_e^2) \propto \frac{1}{\sigma_e^2}, \quad \epsilon \leq \rho \leq 1 - \epsilon,\ \sigma_e^2 > 0, \boldsymbol{\beta} \in \mathcal{R}^p,
\end{aligned}$$

for $\epsilon > 0$ small.

The posterior distribution can be obtained in terms of posterior conditionals using the chain rule of probability as follows. First note that, under the HB approach, the random effects $\mathbf{v} = (v_1, \ldots, v_m)'$ are regarded as additional parameters. Then, the joint posterior pdf of the vector of parameters $\boldsymbol{\theta} = (\mathbf{v}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$ given the sample values $\mathbf{y}_s$ is given by

$$\pi(\mathbf{v}, \boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s) \quad = \quad \pi_1(\mathbf{v}|\boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s)\pi_2(\boldsymbol{\beta}|\sigma_e^2, \rho, \mathbf{y}_s)\pi_3(\sigma_e^2|\rho, \mathbf{y}_s)\pi_4(\rho|\mathbf{y}_s), \tag{2.18}$$

where the conditional pdfs $\pi_1, \ldots, \pi_3$ have known forms, but not π_4 (see Appendix A.2). However, since ρ is in a closed interval from $(0, 1)$, we can generate values from π_4 using a grid method, for more details see Molina et al. (2014). Samples from $\boldsymbol{\theta} = (\mathbf{v}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$ can then be generated directly from the posterior distribution in (2.18), avoiding the use of Markov Chain Monte Carlo (MCMC) methods. Under general conditions, a proper posterior distribution is guaranteed, see Molina et al. (2014).

Given $\boldsymbol{\theta}$, population variables Y_{ij} are all independent, satisfying

$$Y_{ij}|\boldsymbol{\theta} \overset{ind}{\sim} N(\mathbf{x}_{ij}'\boldsymbol{\beta} + v_i, \sigma_e^2 k_{ij}^2), \quad j = 1, \ldots, N_i,\ i = 1, \ldots, m. \tag{2.19}$$

Consider the decomposition of the domain vector $\mathbf{y}_i = (Y_{i1}, \ldots, Y_{iN_i})'$ in terms of sample and out-of-sample elements $\mathbf{y}_i = (\mathbf{y}_{is}', \mathbf{y}_{ir}')'$. The posterior predictive pdf of $\mathbf{y}_{ir}$ is then given by

$$f(\mathbf{y}_{ir}|\mathbf{y}_s) = \int \prod_{j \in r_i} f(Y_{ij}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}_s)d\boldsymbol{\theta}.$$

Finally, the HB estimator of a domain parameter $H_i = h(\mathbf{y}_i)$ is given by

$$\hat{H}_i^{HB} = E_{\mathbf{y}_{ir}}(H_i|\mathbf{y}_s) = \int h(\mathbf{y}_i) f(\mathbf{y}_{ir}|\mathbf{y}_s) d\mathbf{y}_{ir}. \tag{2.20}$$

The HB estimator can be approximated by Monte Carlo. For this, we first generate samples from the posterior $\pi(\boldsymbol{\theta}|\mathbf{y}_s)$ given in (2.18). We generate a value $\rho^{(a)}$ from

$\pi_4(\rho|\mathbf{y}_s)$ using a grid method; then, a value $\sigma_e^{2(a)}$ is generated from $\pi_3(\sigma_e^2|\rho^{(a)}, \mathbf{y}_s)$; next $\boldsymbol{\beta}^{(a)}$ is generated from $\pi_2(\boldsymbol{\beta}|\sigma_e^{2(a)}, \rho^{(a)}, \mathbf{y}_s)$ and, finally, $\mathbf{v}^{(a)}$ is generated from $\pi_1(\mathbf{v}|\boldsymbol{\beta}^{(a)}, \sigma_e^{2(a)}, \rho^{(a)}, \mathbf{y}_s)$. This process is repeated a large number A of times to get a random sample $\boldsymbol{\theta}^{(a)}$, $a = 1, \ldots, A$ from $\pi(\boldsymbol{\theta}|\mathbf{y}_s)$. Now for each generated value $\boldsymbol{\theta}^{(a)}$ from $\pi(\boldsymbol{\theta}|\mathbf{y}_s)$, we generate the out-of-sample values $\{Y_{ij}^{(a)}, j \in r_i\}$ from the distribution defined in (2.19). Thus, for each domain i, we have generated an out-of-sample vector $\mathbf{y}_{ir}^{(a)} = \{Y_{ij}^{(a)}, j \in r_i\}$, and we have also the available sample data $\mathbf{y}_{is}$. Putting them together, we construct the full population vector $\mathbf{y}_i^{(a)} = (\mathbf{y}_{is}', (\mathbf{y}_{ir}^{(a)})')'$. Now using $\mathbf{y}_i^{(a)}$, we compute the domain parameter $H_i^{(a)} = h(\mathbf{y}_i^{(a)})$. In the particular case of estimating a FGT poverty indicator, we have $H_i = F_{\alpha i} = h_\alpha(\mathbf{y}_i)$ given in (2.17). Then, in Monte Carlo replicate a, we calculate $F_{\alpha i}^{(a)} = h_\alpha(\mathbf{y}_i^{(a)})$. Finally, the HB estimator is approximated as

$$\hat{F}_{\alpha i}^{HB} \approx \frac{1}{A} \sum_{a=1}^{A} F_{\alpha i}^{(a)}. \tag{2.21}$$

Advantages and deficiencies of HB method are listed below.

Advantages:

- It is based on unit level data, which are richer than area level data and uses much larger sample size to fit the model.
- HB method can be applied to estimate general indicators defined as function of the model response variables Y_{ij}.
- HB estimators are model-unbiased.
- HB estimators are optimal in terms of minimizing the posterior variance.
- EB and HB methods are expected to give practically the same point estimates, see Molina et al. (2014). Thus, the proposed HB method has good frequentist properties.
- Once the model is fitted, estimates can be obtained at whatever subdomain level.
- The proposed HB approach does not require the use of MCMC methods and therefore avoids the need of monitoring the convergence of Monte Carlo chains.
- Bootstrap methods for MSE estimation are not needed. Therefore, total computational time is considerably lower than in EB method.
- Calculation of credible intervals or other posterior summaries are straightforward.

Disadvantages:

- It is based on model assumptions. Hence, model checking is crucial.

- HB estimators are not design-unbiased and can be seriously biased under informative sampling.
- HB estimators can be severely affected by unit level outliers.
- HB method is not directly extendable to more complex models without losing some of the mentioned advantages like avoiding MCMC.

2.3 Simulation studies

This section illustrates some of the mentioned advantages and drawbacks of the considered poverty mapping methods through simulation studies. Concretely, we will report results of simulations under three different scenarios: (i) Nested error model with simple random sampling. (ii) Nested error model with informative sampling. (iii) Nested error model with outliers.

Simulations were implemented in the statistical software environment R (R development core team 2013) using the package `lme4` (Bates et al., 2014), which fits Gaussian linear and nonlinear mixed-effects models, and the package `sae` (Molina and Marhuenda, 2015), which contains functions for small area estimation, including calculation of direct, FH and EB estimates along with their model MSE estimates.

2.3.1 Nested error model with simple random sampling

We consider the same model-based simulation setup as in Molina et al. (2014), where data are generated at the unit level following the nested error model (2.9). However, here we will also include FH estimators derived from the FH area level model obtained using the domain means of the auxiliary variables as covariates. In addition, we include ELL and Census EB estimators. The population is composed of $N = 20,000$ units, distributed in $m = 80$ domains with $N_i = 250$ units in each domain. We consider two auxiliary variables X_1 and X_2 with known values for all the population units. Their values are generated as $x_{q,ij} \sim \text{Bern}(p_{qi})$, $q = 1, 2$, with success probabilities $p_{1i} = 0.3 + 0.5i/m$ and $p_{2i} = 0.2$, $i = 1, \ldots, m$. Response variables Y_{ij} are generated from the nested error model (2.9) and the target variables are $E_{ij} = \exp(Y_{ij})$. The true values of the regression coefficients are $\boldsymbol{\beta} = (3, 0.03, -0.04)'$. Variances of domain effects and errors are taken as $\sigma_v^2 = 0.15^2$ and $\sigma_e^2 = 0.5^2$ respectively. The poverty line is set to $z = 12$, which is approximately 0.6 times the median of $\{E_{ij}; j = 1, \ldots, N_i,\ i = 1, \ldots, m\}$ for a population generated as described before, which is the official definition of poverty line used in EU countries. We draw a sample s_i of size $n_i = 50$, $i = 1, \ldots, m$, using sample random sampling without replacement (srswor), independently from each domain i.

A total of $K = 1,000$ population vectors $\mathbf{y}^{(k)}$, $k = 1, \ldots, K$, were generated from the nested error model (2.9) with the mentioned values of model parameters and auxiliary variables. For each Monte Carlo population $k = 1, \ldots, K$, we calculated the true domain poverty incidences and poverty gaps. Then, we selected the sample s, which is kept fixed across Monte Carlo replicates. Using the sample data $\{(Y_{ij}, x_{1,ij}, x_{2,ij}); j \in s_i, i = 1, \ldots, m\}$ and the population data on the auxiliary variables, we computed direct estimates, FH, ELL, EB, Census EB and HB estimates of poverty incidence ($\alpha = 0$) and poverty gap ($\alpha = 1$) for each domain $i = 1, \ldots, m$. FH, ELL and EB estimates were obtained using REML fitting method. The FH model (2.6) is fitted using as area level covariates the domain means of the two considered auxiliary variables, that is, $\mathbf{x}_i = (1, \bar{X}_{1,i}, \bar{X}_{2,i})'$, where $\bar{X}_{q,i} = N_i^{-1} \sum_{j=1}^{N_i} x_{q,ij}$, $q = 1, 2$.

For the Monte Carlo population k, let $F_{\alpha i}^{(k)}$ be the true poverty indicator for domain i and $\hat{F}_{\alpha i}^{(k)}$ be one of the estimates (direct (DIR), FH, ELL, EB, Census EB or HB). Relative bias (RB) and relative root MSE (RRMSE) of an estimator $\hat{F}_{\alpha i}$ under model (2.9) are approximated empirically as

$$\text{RB}_m(\hat{F}_{\alpha i}) = \frac{K^{-1} \sum_{k=1}^{K} \left(\hat{F}_{\alpha i}^{(k)} - F_{\alpha i}^{(k)} \right)}{K^{-1} \sum_{k=1}^{K} F_{\alpha i}^{(k)}}, \quad \text{RRMSE}_m(\hat{F}_{\alpha i}) = \frac{\sqrt{K^{-1} \sum_{k=1}^{K} \left(\hat{F}_{\alpha i}^{(k)} - F_{\alpha i}^{(k)} \right)^2}}{K^{-1} \sum_{k=1}^{K} F_{\alpha i}^{(k)}}.$$

For each estimator $\hat{F}_{\alpha i}$, the absolute RB (ARB) and the RRMSE are averaged across domains as

$$\overline{\text{ARB}}_\alpha = m^{-1} \sum_{i=1}^{m} |\text{RB}_m(\hat{F}_{\alpha i})|, \quad \overline{\text{RRMSE}}_\alpha = m^{-1} \sum_{i=1}^{m} \text{RRMSE}_m(\hat{F}_{\alpha i}).$$

Figure 2.1 depicts the percent values of RB (left) and RRMSE (right) of the estimators of the domain poverty gaps F_{1i} for each domain i. EB and Census EB estimates are not shown in these plots because they are both practically equal to HB estimates and are plotted separately in Figure 2.2. Figure 2.1 left shows that direct, ELL and HB estimators are practically unbiased. In contrast, FH estimators display a substantial negative bias. Concerning efficiency, Figure 2.1 right shows that HB estimators have the smallest RRMSE whereas ELL estimators are the ones with the largest RRMSE. Conclusions for the poverty incidence F_{0i} are very similar.

Table 2.1 displays averages across domains of ARB and RRMSE of all the estimators, for both poverty incidence and poverty gap. We see that FH estimator exhibits a large ARB (over 6% for poverty incidence and close to 15% for poverty gap), whereas EB, HB and Census HB estimators have a very small ARB ($< 1\%$). The latter estimators also achieve the smallest RRMSEs (slightly over 20% for poverty incidence and over 25% for

poverty gap). The largest RRMSE is obtained by ELL estimator (over 58%). Note that both ARB and RRMSE increase when estimating the poverty gap, because the poverty gap depends to a greater extent on the extreme of the left tail of the income distribution, which is more difficult to estimate correctly from a (finite) sample.

These results indicate that HB estimators are practically unbiased and clearly the most efficient among the considered estimators when the nested error model holds and the sample is drawn with srswor within each domain. The bias of FH estimators is due to the fact that they are attaching most of the weight to the regression-synthetic component, which relies exactly on the model, but here data Y_{ij} are generated from the unit level model (2.9) and the domain means of the covariates $\bar{X}_{q,i} = N_i^{-1} \sum_{j=1}^{N_i} x_{q,ij}$ are not linearly related with the poverty indicators $F_{\alpha i}$. Thus, FH model fails due to non-linearity of the poverty indicators $F_{\alpha i}$ in the domain level covariates $\bar{X}_{q,i}$, $k = 1, 2$, even if the unit level model holds exactly, because of the non-linearity of $F_{\alpha i}$ as function of Y_{ij}.

Figure 2.1: Percent RB (left) and RRMSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1i} for each domain i under the nested error model with srswor.

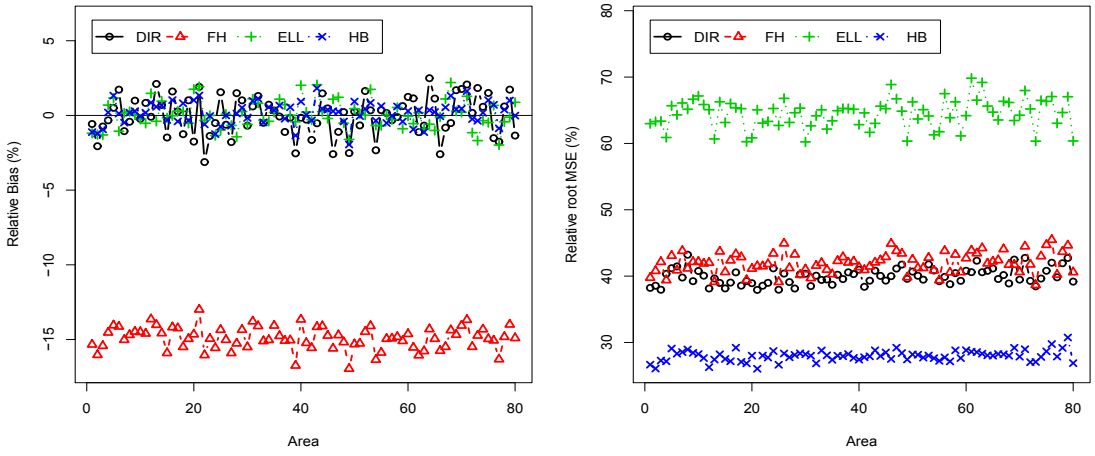


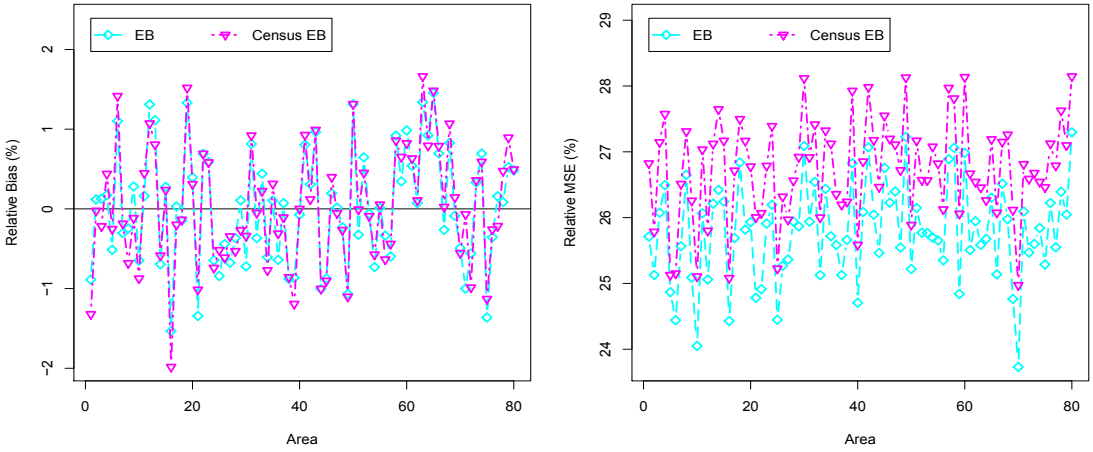
Figure 2.2 depicts percent RB (left) and RRMSE (right) of EB and Census EB estimates of the poverty gap F_{1i} for each domain i. This figure shows the great similarity of EB and Census EB estimates of F_{1i}, even if sampling fractions in this simulation study are not so small ($n_i/N_i = 1/5$, $i = 1, \ldots, m$).

Next, we study ELL estimator of the MSE of $\hat{F}_{\alpha i}^{ELL}$ given in (2.10). Figure 2.3 depicts the true model MSE of ELL estimators of the poverty gap F_{1i}, labeled "True MSE ELL" and the means across 10,000 simulations of ELL estimates of the MSE, $\text{MSE}_m(\hat{F}_{\alpha i}^{ELL})$, labeled "MSE ELL", for each domain i. This figure shows that ELL estimates of the model MSE do not really track the true model MSEs for each domain even if we have

Method	Average ARB (%)		Average RRMSE (%)	
	F_{0i}	F_{1i}	F_{0i}	F_{1i}
Direct	0.99	1.26	28.53	36.33
FH	6.34	14.78	26.26	38.16
HB	0.48	0.65	20.15	25.43
EB	0.51	0.67	20.41	25.75
Census EB	0.55	0.69	21.15	26.71
ELL	1.31	1.69	47.39	58.63

Table 2.1: Averages across domains of percent ARB and RRMSE for direct, FH, HB, EB, Census EB and ELL estimators of poverty incidence F_{0i} and poverty gap F_{1i}, under the nested error model with srswor.

Figure 2.2: Percent RB (left) and RRMSE (right) of EB and Census EB estimators of poverty gap F_{1i} for each domain i under the nested error model with srswor.



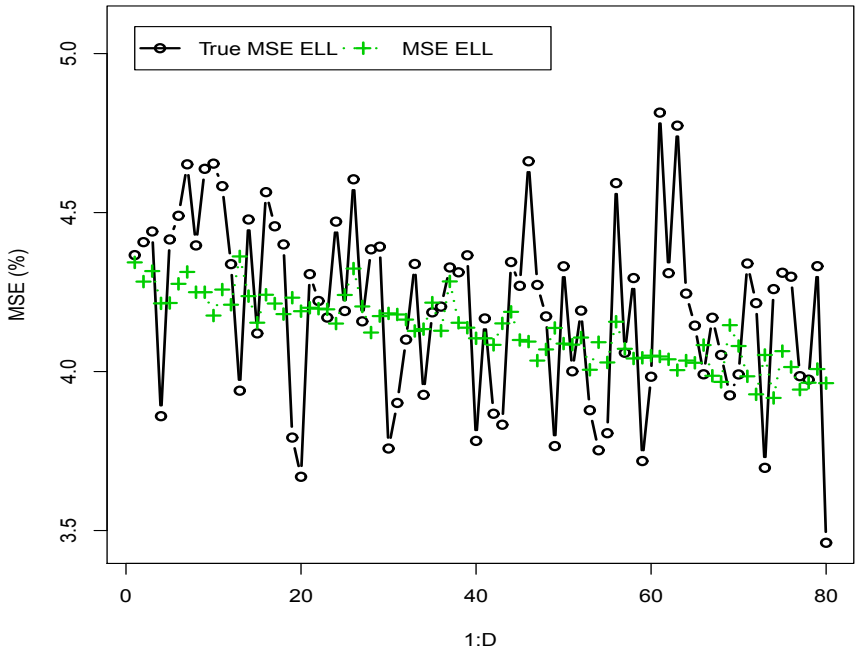
considered here random effects for the domains in the model (i.e., sampling clusters equal to domains). In the case that clusters are different from the domains, if we consider the original ELL method that includes only cluster effects but area effects are significant, then ELL estimates might seriously underestimate the MSE.

For the EB estimator, the parametric bootstrap procedure proposed by Molina and Rao (2010) approximates the true MSE reasonably well, see Molina and Rao (2010). For HB estimator, posterior variance, approximated by Monte Carlo, is taken as measure of uncertainty.

2.3.2 Nested error model with informative sampling

We consider the same setup as in the previous simulation study, with the same population sizes, model parameters, auxiliary variables and poverty line. The only difference is that, in this simulation study, samples are drawn with informative sampling. When

Figure 2.3: True MSE of ELL estimators of poverty gap F_{1i} and mean across simulations of ELL estimator of the MSE for each domain i, under the nested error model with srswor.



the sampling is informative, the probability of a sample depends on the values of the population vector $\mathbf{y}$. Thus, under this setup, the simulations need to be performed with respect to the joint distribution of $(\mathbf{y}, s)$; that is, in each Monte Carlo replicate k, we draw a population vector $\mathbf{y}^{(k)}$ and, given $\mathbf{y}^{(k)}$, we draw a sample $s^{(k)}$. A total of $K = 1,000$ population vectors $\mathbf{y}^{(k)}$, $k = 1, \ldots, K$, are generated from the true nested error model (2.9). Again, we consider that the target variables are $E_{ij} = \exp(Y_{ij})$. The sample $s^{(k)}$ is drawn by Poisson sampling, with inclusion probabilities π_{ij} depending on a random variable Z_{ij} that is correlated with the unexplained part of Y_{ij}, that is, the model errors e_{ij}. Thus, for each population unit j from domain i, we generate a Bernoulli random value $Q_{ij} \sim \text{Bern}(\pi_{ij})$, with $\pi_{ij} = b^{-1}\exp(-aZ_{ij})$, where $a > 0$, $b > 0$ and $Z_{ij} \sim \text{Gamma}(\tau_{ij}, \theta_{ij})$. To choose the values of τ_{ij} and θ_{ij}, we consider two cases: low and high level of informativeness. In the first case, we take $\tau_{ij} = 5(3 + 0.1e_{ij})$ and $\theta_{ij} = 0.25(3 + 0.1e_{ij})$, which yield random values Z_{ij} with a 20% correlation with the model errors e_{ij}. In the second case, we take $\tau_{ij} = 5(4.5 + 1.5e_{ij})$ and $\theta_{ij} = 0.25(4.5 + 1.5e_{ij})$, yielding Z_{ij} with a 80% correlation with e_{ij}, which represents a high level of informativeness. Note that, under this set up, the sample size is random because each unit in the population comes to the sample depending on its random value Q_{ij}. To make this simulation study comparable with the one in previous section,

we wish to have a similar average domain sample size as before. This is achieved approximately by considering $a = 0.05$ and $b = 2.5$ when the informativeness level is low and taking $a = 0.02$ and $b = 4$ when the informative level is high. With the sample $s^{(k)}$ from each population, we compute the five estimators, namely direct, FH, EB, ELL and HB estimators. We excluded here Census EB estimators because of their great similarity with EB estimators.

Figure 2.4 plots RBs (left) and RRMSEs (right) under the model (2.9) and the design of the estimators of the poverty gap F_{1i} when the informativeness level is low. Again, EB estimator is excluded because it provides nearly the same results as HB. For low level of informativeness, Figure 2.4 left shows that the negative bias of FH estimates, observed in the simulation with srswor, still persists, while the rest of the estimators are almost unbiased. HB estimator still displays the smallest RRMSE, and ELL estimator performs the worst in terms of RRMSE. For the poverty incidence F_{0i}, conclusions are similar. These conclusions are confirmed by the averages across domains shown in Table 2.2 for both poverty incidence and poverty gap. On average, the direct estimator has the smallest ARB (about 0.7% for poverty incidence and 0.9% for poverty gap), followed by EB and HB estimators with a bias below 1.4% for both poverty incidence and gap, the smallest RRMSE is for EB estimator (less than 21% for poverty incidence and than 26% for poverty gap) and the largest for ELL estimator (over 47% for poverty incidence and over 58% for poverty gap).

Figure 2.5 plots RB (left) and RRMSE (right) of the estimators of the poverty gap F_{1i} when the level of informativeness is high. In this case, Figure 2.5 left shows a negative bias for the FH estimator and a large positive bias of HB and ELL estimators. Looking at Figure 2.5 right, it appears that now, direct and FH estimates, which are calculated using the true inclusion probabilities, exhibit the smallest RRMSE. Again, conclusions are similar for the poverty incidence F_{0i}. Table 2.3 lists the averages across domains of ARB and RRMSE for all the considered estimators of the poverty incidence and poverty gap. In this case, the direct estimator has the smallest average absolute relative bias (about 0.6% for poverty gap), whereas the average ARB is the largest for ELL estimator (97.3%).

To summarize, EB and HB methods are only mildly affected under low level of informativeness, measured in terms of correlation between the design variable used in the inclusion probabilities and the response variable. When the degree of informativeness is high, these two methods are certainly affected because they do not take into account the sampling design. The effect of informative sampling on FH estimator seems to be smaller, and its negative bias is again due to a non-linearity problem of FH model because data actually follows the nested error linear regression

model for log income at the unit level. In Chapter 3, we propose a method to handle informative sampling in the case of unit level models.

Method	Average ARB (%)		Average RRMSE (%)	
	F_{0i}	F_{1i}	F_{0i}	F_{1i}
Direct	0.74	0.91	71.69	38.92
FH	10.47	19.26	30.33	43.38
HB	1.10	1.38	20.29	35.63
EB	1.04	1.25	20.48	25.86
ELL	1.63	1.98	47.39	58.65

Table 2.2: Averages across domains of percent ARB and RRMSE for direct, FH, HB, EB and ELL estimators of incidence F_{0i} and poverty gap F_{1i}, under low informativeness.

Method	Average ARB (%)		Average RRMSE (%)	
	F_{0i}	F_{1i}	F_{0i}	F_{1i}
Direct	0.59	0.65	23.62	25.69
FH	6.94	9.21	23.83	29.40
HB	61.64	76.95	66.05	84.95
EB	61.60	73.68	66.08	84.89
ELL	61.69	76.98	72.94	97.29

Table 2.3: Averages across domains of percent ARB and RRMSE for direct, FH, HB, EB and ELL estimators of incidence F_{0i} and poverty gap F_{1i}, under high informativeness.

2.3.3 Nested error model with outliers

In this section, we conduct a simulation study under exactly the same conditions as in Section 2.3.1, but generating the model errors e_{ij} from a mixture of normal distributions with different variances in order to create outliers. Concretely, in this simulation study, we generate model errors as $e_{ij} \sim (1-\varepsilon)\, N(0, \sigma_e^2) + \varepsilon\, N(0, R\sigma_e^2)$, where ε is generated as $\varepsilon \sim \text{Bern}(p)$. We consider two fractions of outliers, $p = 0.01$ and $p = 0.05$, and two values for the factor R in the variance of the outliers, namely $R = 10$ and $R = 100$.

Using the above mechanism to generate model errors, a total of $K = 1,000$ population vectors $\mathbf{y}^{(k)}, k = 1, \ldots, K$, were generated from the nested error model (2.9). Then, we calculated true domain poverty incidences and gaps. Note that the outliers considered in this simulation study are not recording errors in the sample data. They are actual representative outliers appearing in the population. Thus, they are actually realizations of the distribution with heavier tails obtained from the normal mixture, and true values of poverty indicators must include the outliers generated in the population.

The sample is drawn by srswor within each domain as in Section 2.3.1, keeping the set of sample units s fixed across simulations. With each Monte Carlo sample, direct,

Figure 2.4: Percent RB (left) and RRMSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1i} for each domain i under low informativeness.

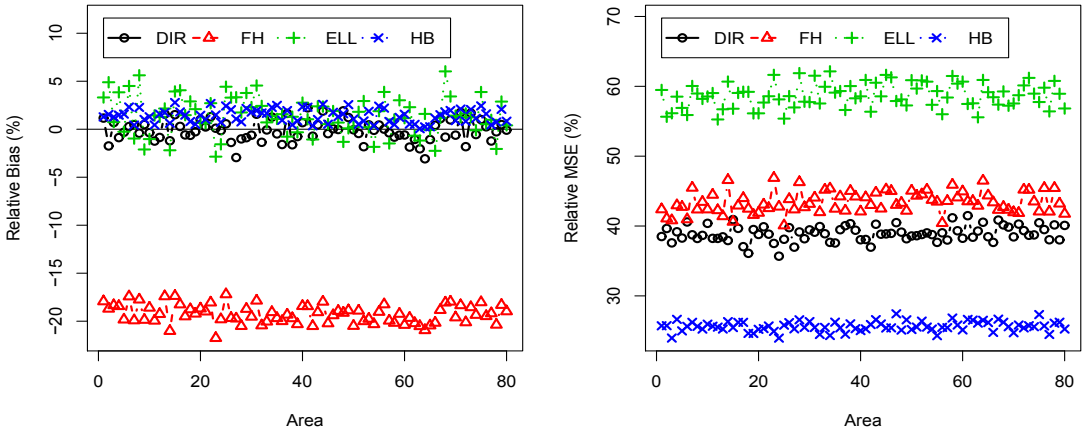
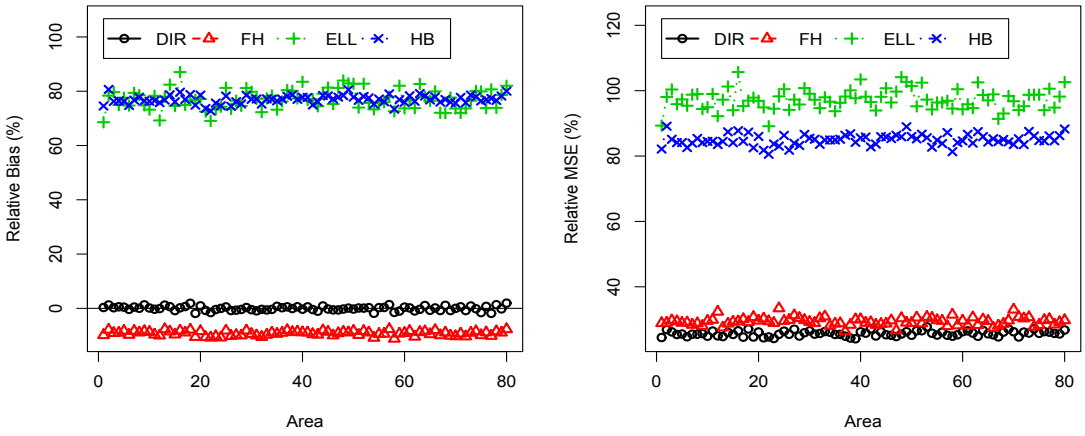


Figure 2.5: Percent RB (left) and RRMSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1i} for each domain i under high informativeness.



FH, EB, ELL and HB estimators were computed.

We report here results for the cases of less frequent mild outliers ($p = 0.01$ and $R = 10$), and of more frequent and extreme outliers ($p = 0.05$ and $R = 100$). For the first case, results for the poverty gap are plotted in Figure 2.6. Again, EB is excluded in the plots because it provides similar results as HB. Figure 2.6 left and right show that direct estimators are not practically affected by the outliers, which is expected because this estimator does not rely on any model assumption. Similarly, FH estimator is less affected by the outliers because the observed negative bias is again due to non-linearity problems. HB and ELL estimators show a moderate bias, but still HB estimator achieves the lowest RRMSE. Averages across domains of ARB and RRMSE for all estimators of poverty incidence and poverty gap are shown in Table 2.4. The ARB of EB and HB estimators is small (around 4% for pov. incidence and 5% for pov. gap), and the RRMSE has increased only about 0.5% with respect to the case of no outliers (see Table 2.2) and it is still acceptable (around 21% for pov. incidence and 26% for pov. gap).

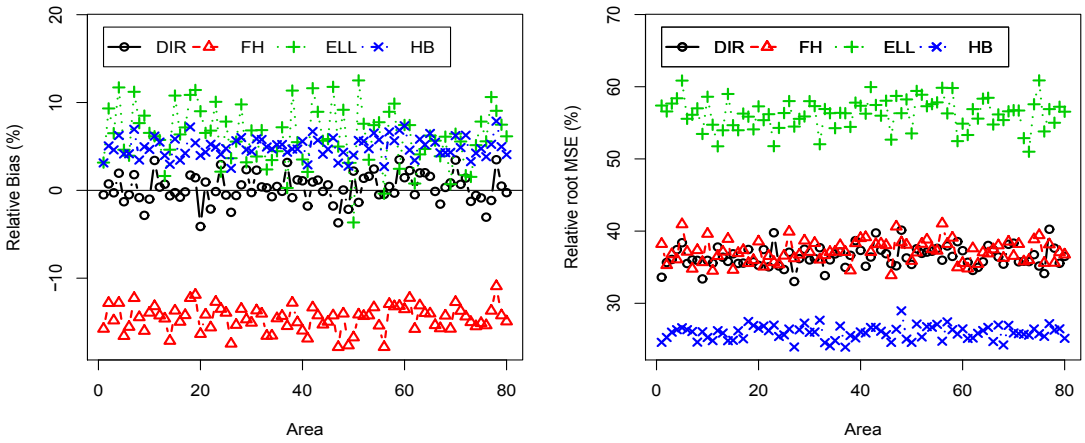
For the case of more frequent and extreme outliers ($p = 0.05$ and $R = 100$), Figure 2.7 left shows that, in this case, HB and, to a greater extent, ELL, display a very large positive RB, see also Table 2.5 reporting averages across domains. Note that the RRMSE of ELL estimator reaches 226.63% for the poverty gap. In this simulation study, FH estimates perform better than in the previous simulation studies, and this could be due to the fact that, since FH model is less correct when outliers are present, the FH estimator is attaching more weight to the direct estimator, which is practically unbiased. EB, HB and ELL estimators are severely biased when data contains frequent extreme outliers, performing even worse than under high level of informative sampling, but are not too much affected under rare and not so extreme outliers. These methods are based on model assumptions and are not robust to strong model misspecification when the true error distribution has very heavy tails as in the mixture model considered here with $p = 0.05$ and $R = 100$.

We plan to explore estimation methods for complex parameters that are robust to outliers. Note that previous work on robust M-estimation, e.g., Sinha and Rao (2009), focused on estimating domain means only. Small area estimation methods for poverty mapping based on robust M-quantile models have been proposed by Tzavidis et al. (2008).

2.4 Application Spanish SILC data

We apply the methods presented above to real data from the Spanish Survey on Income and Living Conditions (SILC) of year 2006. The EB procedure has been excluded in this

Figure 2.6: Percent RB (left) and RRMSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1i} for each domain i under nested error model with outliers ($p = 0.01$ and $R = 10$).



Method	Average ARB (%)		Average RRMSE (%)	
	F_{0i}	F_{1i}	F_{0i}	F_{1i}
Direct	0.92	1.18	28.54	36.82
FH	6.16	14.67	26.10	37.55
HB	3.95	4.95	20.81	26.22
EB	3.88	4.79	20.99	26.42
ELL	4.93	6.14	46.65	56.52

Table 2.4: Averages across domains of percent ARB and RRMSE for direct, FH, HB, EB and ELL estimators of incidence F_{0i} and poverty gap F_{1i}, under under nested error model with outliers ($p = 0.01$ and $R = 10$).

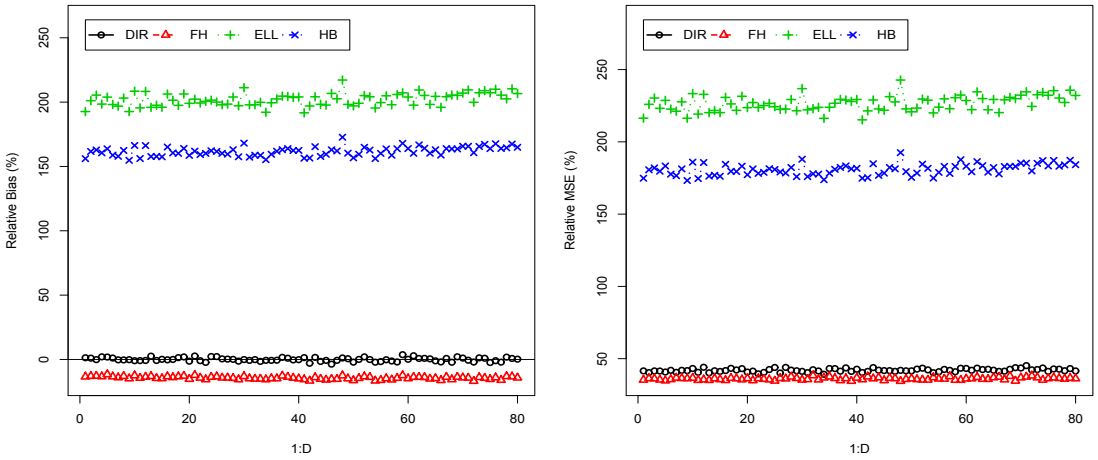
application because it is computationally less efficient than the HB but provides roughly the same point estimates than the HB method, see Molina et al. (2014).

The SILC collects microdata on income, poverty and social exclusion and living conditions, in a timely and comparable way across European Union (EU) countries. The results obtained from the SILC are used for the structural index of social cohesion. This survey provides reliable estimates for the overall Spain and for large Spanish regions (Autonomus Comunities) but it does not allow estimation for Spanish provinces because of the small SILC sample sizes in some of them.

In this application, we estimate poverty indicators in the $m = 52$ provinces by gender. The overall sample size is 17,739 for women and 16,650 for men. The population size is 22,077,565 for women and 21,509,962 for men.

As auxiliary variables, we included the same as in Molina et al. (2014), namely the indicators of quinquennial age groups, of having Spanish nationality, of the three levels of the variable education level and of the three categories of the variable labor force

Figure 2.7: Percent RB (left) and RRMSE (right) of direct, FH, HB and ELL estimators of poverty gap F_{1i} for each domain i under under nested error model with outliers ($p = 0.05$ and $R = 100$)



Method	Average ARB (%)		Average RRMSE (%)	
	F_{0i}	F_{1i}	F_{0i}	F_{1i}
Direct	0.96	1.20	29.68	41.99
FH	5.66	14.33	26.65	36.10
HB	74.13	161.73	86.87	180.88
EB	74.11	161.59	86.95	180.81
ELL	92.64	201.97	111.32	226.63

Table 2.5: Averages across domains of percent ARB and RRMSE for direct, FH, HB, EB and ELL estimators of incidence F_{0i} and poverty gap F_{1i}, under under nested error model with outliers ($p = 0.05$ and $R = 100$).

status. Similarly as in Molina et al. (2014), full census matrices $\mathbf{X}_i$ were constructed by replicating each record in the Spanish Labor Force Survey (LFS) a number of times equal to its LFS sampling weight. These matrices $\mathbf{X}_i$ are treated as the census matrices because the LFS has a very large sample size.

The welfare measure E_{ij} in the SILC is the equivalent annual net income, which is defined as the household annual net income divided by a measure of household size calculated according to the scale defined by OCDE. The poverty line was also computed according with this welfare measure, as $z = 0.6\times$Median(welfare). Finally, due to the right skewness of the equivalent annual net income, we also consider the same transformation as in Molina et al. (2014), $Y_{ij} = T(E_{ij}) = \log(E_{ij} + c)$ where c is selected such that the residuals obtained from the model fit are approximately symmetric. Here, we report the resulting direct, FH and HB estimates together with their estimated coefficients of variation (or estimated RRMSEs) under the model. For the HB method, we considered a grid of $R = 1,000$ values of ρ and $A = 1,000$ Monte

Carlo replicates.

Province	Dom	n_i	$\hat{F}_{0i}^{DIR}$	$\hat{F}_{0i}^{FH}$	$\hat{F}_{0i}^{HB}$	$cv_m(\hat{F}_{0i}^{DIR})$	$cv_m(\hat{F}_{0i}^{FH})$	$cv_m(\hat{F}_{0i}^{HB})$
Soria	42	17	22.27	20.92	33.75	42.70	25.37	18.03
Gerona	17	138	5.08	6.15	17.06	34.86	27.84	10.47
Ciudad Real	13	239	33.30	31.73	30.54	12.52	10.62	6.70
Sevilla	41	491	21.69	22.73	24.33	10.05	8.99	5.37
Barcelona	8	1483	11.40	11.41	13.80	7.86	7.77	3.92

Table 2.6: Results for poverty incidences for women: Direct, FH and HB estimates together with coefficients of variation (%) for Spanish provinces with sample sizes closest to minimun, 0.25, 0.5, 0.75 quantiles and maximum.

Province	Dom	n_i	$\hat{F}_{0i}^{DIR}$	$\hat{F}_{0i}^{FH}$	$\hat{F}_{0i}^{HB}$	$cv_m(\hat{F}_{0i}^{DIR})$	$cv_m(\hat{F}_{0i}^{FH})$	$cv_m(\hat{F}_{0i}^{HB})$
Soria	42	24	13.87	14.06	24.32	51.87	27.85	18.48
Lérida	25	127	19.08	17.36	24.72	22.47	18.85	10.09
Jaén	23	233	34.01	29.49	28.77	12.38	10.64	7.48
Palmas, Las	35	458	25.85	23.85	24.97	13.52	12.08	5.42
Barcelona	8	1358	8.62	8.77	11.08	9.38	9.07	4.56

Table 2.7: Results for poverty incidences for men: Direct, FH and HB estimates together with coefficients of variation (%) for Spanish provinces with sample sizes closest to minimum, 0.25, 0.5, 0.75 quantiles and maximum.

Province	Dom	n_i	$\hat{F}_{1i}^{DIR}$	$\hat{F}_{1i}^{FH}$	$\hat{F}_{1i}^{HB}$	$cv_m(\hat{F}_{1i}^{DIR})$	$cv_m(\hat{F}_{1i}^{FH})$	$cv_m(\hat{F}_{1i}^{HB})$
Soria	42	17	9.99	6.35	12.84	60.53	11.81	24.58
Gerona	17	138	1.63	1.86	5.44	40.97	10.57	12.89
Ciudad Real	13	239	7.36	7.70	10.94	14.11	3.03	8.87
Sevilla	41	491	4.31	4.57	8.14	12.88	2.39	6.90
Barcelona	8	1483	3.80	3.79	4.11	10.26	3.38	4.90

Table 2.8: Results for poverty gaps for women: Direct, FH and HB estimates together with coefficients of variation (%) for Spanish provinces with sample sizes closest to minimum, 0.25, 0.5, 0.75 quantiles and maximum.

Province	Dom	n_i	$\hat{F}_{1i}^{DIR}$	$\hat{F}_{1i}^{FH}$	$\hat{F}_{1i}^{HB}$	$cv_m(\hat{F}_{1i}^{DIR})$	$cv_m(\hat{F}_{1i}^{FH})$	$cv_m(\hat{F}_{1i}^{HB})$
Soria	42	24	8.59	5.27	8.62	64.15	14.70	23.83
Lérida	25	127	9.96	7.03	8.75	25.15	9.98	12.99
Jaén	23	233	11.56	10.09	10.50	15.05	4.69	9.82
Palmas, Las	35	458	8.81	7.95	8.85	17.91	5.60	6.97
Barcelona	8	1358	3.12	3.15	3.29	11.60	4.07	5.60

Table 2.9: Results for poverty gaps for men: Direct, FH and HB estimates together with coefficients of variation (%) for Spanish provinces with sample sizes closest to minimum, 0.25, 0.5, 0.75 quantiles and maximum.

Table 2.6 summarizes the obtained results when estimating poverty incidences for Spanish provinces in the case of women. Concretely, the table includes the results for the provinces that present the sample sizes closest to minimum, first quartile, median, third quartile and maximum. The largest estimates of poverty incidence are obtained using the HB method except for the province of Ciudad Real. The estimated coefficients of variation (CVs) are larger for the direct estimators, exceeding 30% for Gerona and 40% for Soria. Thus, we consider that direct estimators are not reliable. FH estimators seem to smooth the direct estimates. HB estimators perform better than FH estimators according to their estimated CVs. Table 2.7 includes the results for men. This table shows more clearly how the FH estimator smooths the values of the direct estimators and the CVs are again larger for direct estimators.

Table 2.8 shows the results for poverty gap in the case of women. Again, HB estimates take larger values in all the selected domains. The CVs of direct estimators are substantially larger than those of the other estimators. Table 2.9 reports the results for men. Direct and FH estimates for Soria differ quite a bit in this case. The estimated CV of the direct estimator for Soria exceeds 60%.

Figures 2.8 and 2.9 depict cartograms of estimated percent poverty incidence and poverty gap respectively, in Spanish provinces for women obtained using direct (top left), FH (top right) and HB estimator (bottom left). The three methods indicate that the provinces with larger poverty incidence and poverty gap are those in the south and west of Spain. In fact, all three estimators point out to Ávila, Badajoz, Cuenca, Ciudad Real, Almería and Jaén as provinces with highest poverty incidence. Nevertheless, according to the HB estimator, the number of provinces with poverty incidence greater than 30% is larger than with the other two estimators. From these figures, it is clear that the FH estimator smooths the direct estimates, with less provinces in the darkest colors. For the poverty gap, the three estimates coincide in the provinces of Badajoz, Ciudad Real, Jaén, Granada, Almería, Cádiz, as the ones with the largest values. Again, the FH estimator seems to smooth the direct estimator.

Figure 2.8: Cartograms of estimated percent poverty incidences, $\hat{F}_{0i}$, in Spanish provinces for women obtained with direct (top left), FH (top right) and HB (bottom left) methods.

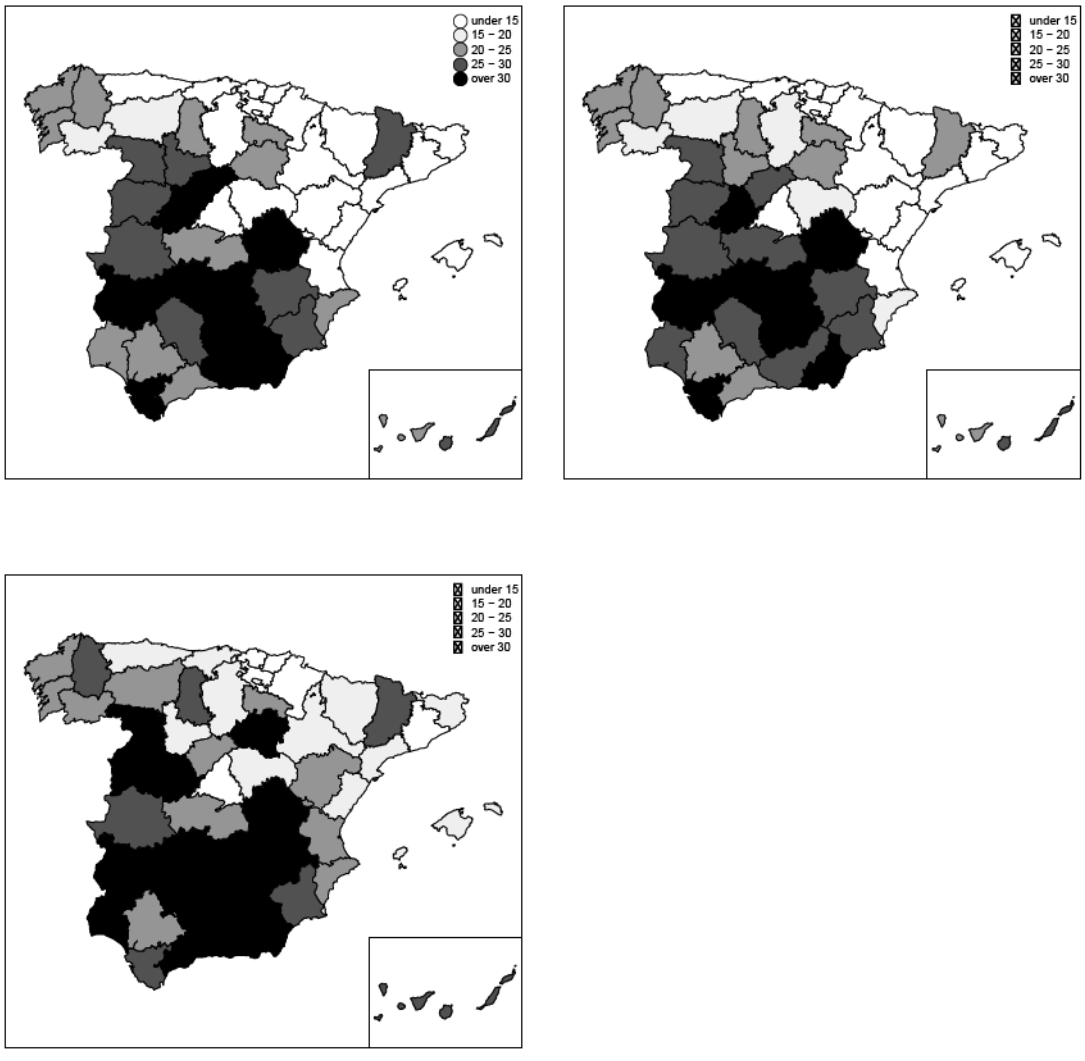


Figure 2.9: Cartograms of estimated percent poverty gaps, $\hat{F}_{1i}$, in Spanish provinces for women obtained with direct (top left), FH (top right) and HB (bottom left) methods.

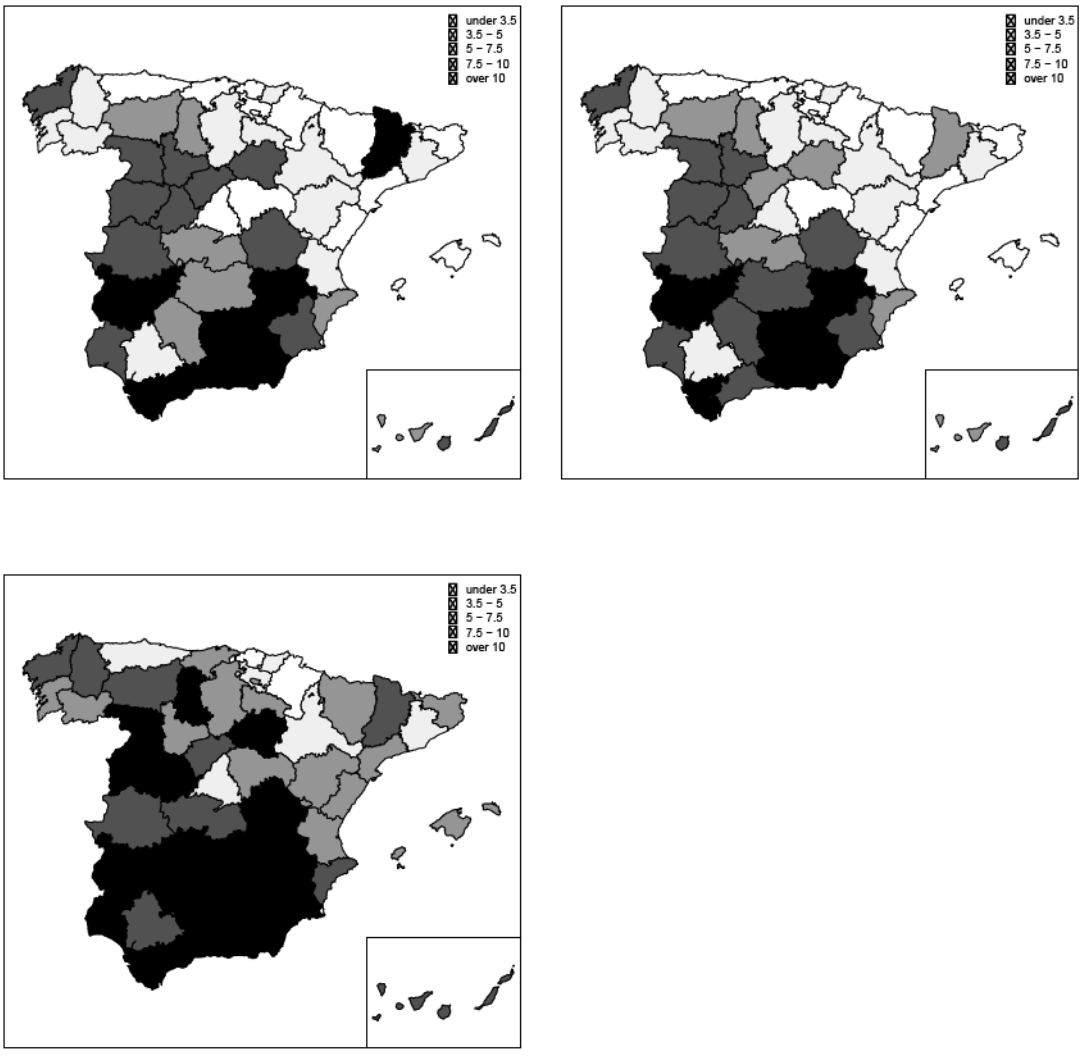


Figure 2.10: Cartograms of estimated percent poverty incidences, $\hat{F}_{0i}$, in Spanish provinces for men obtained with direct (top left), FH (top right) and HB (bottom left) methods.

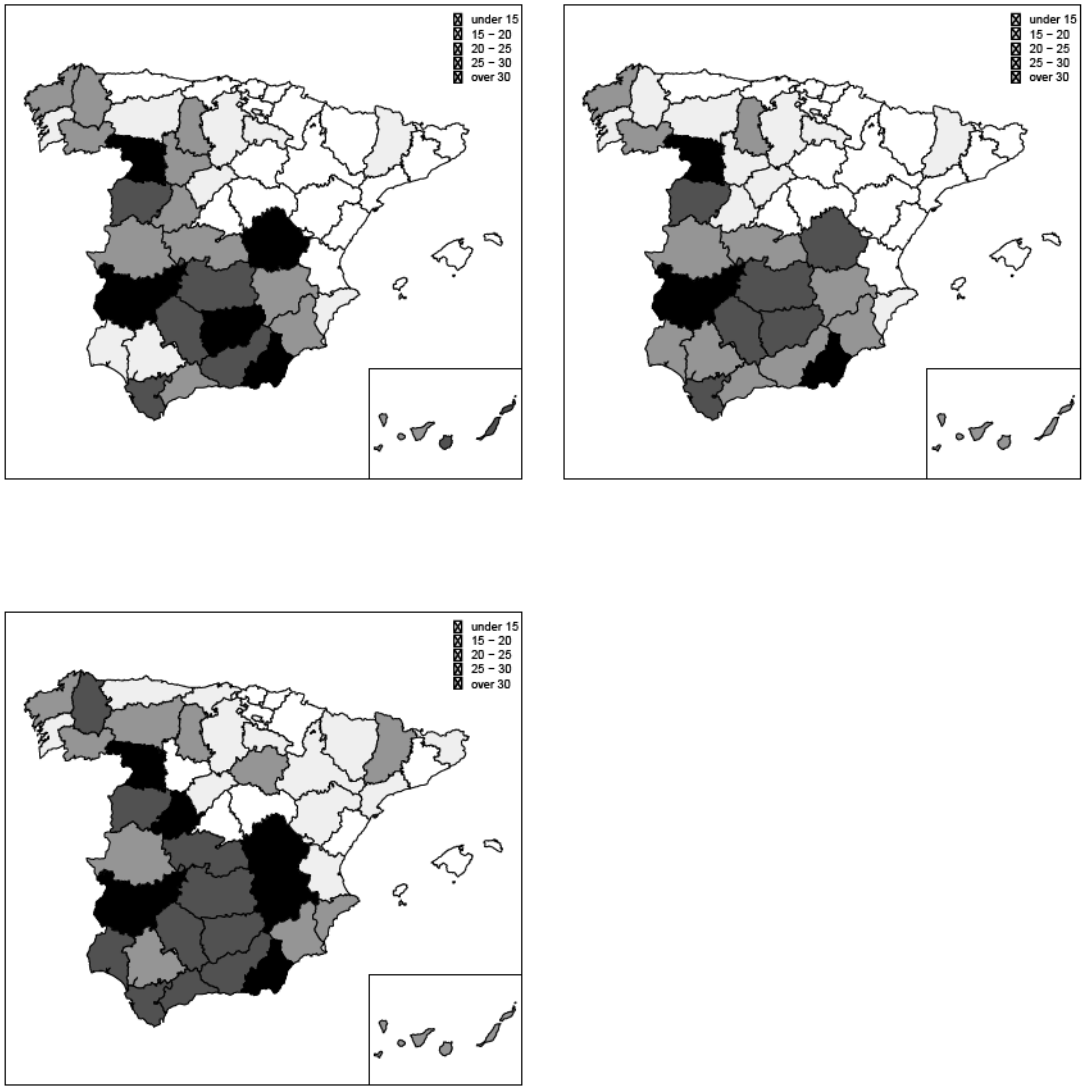
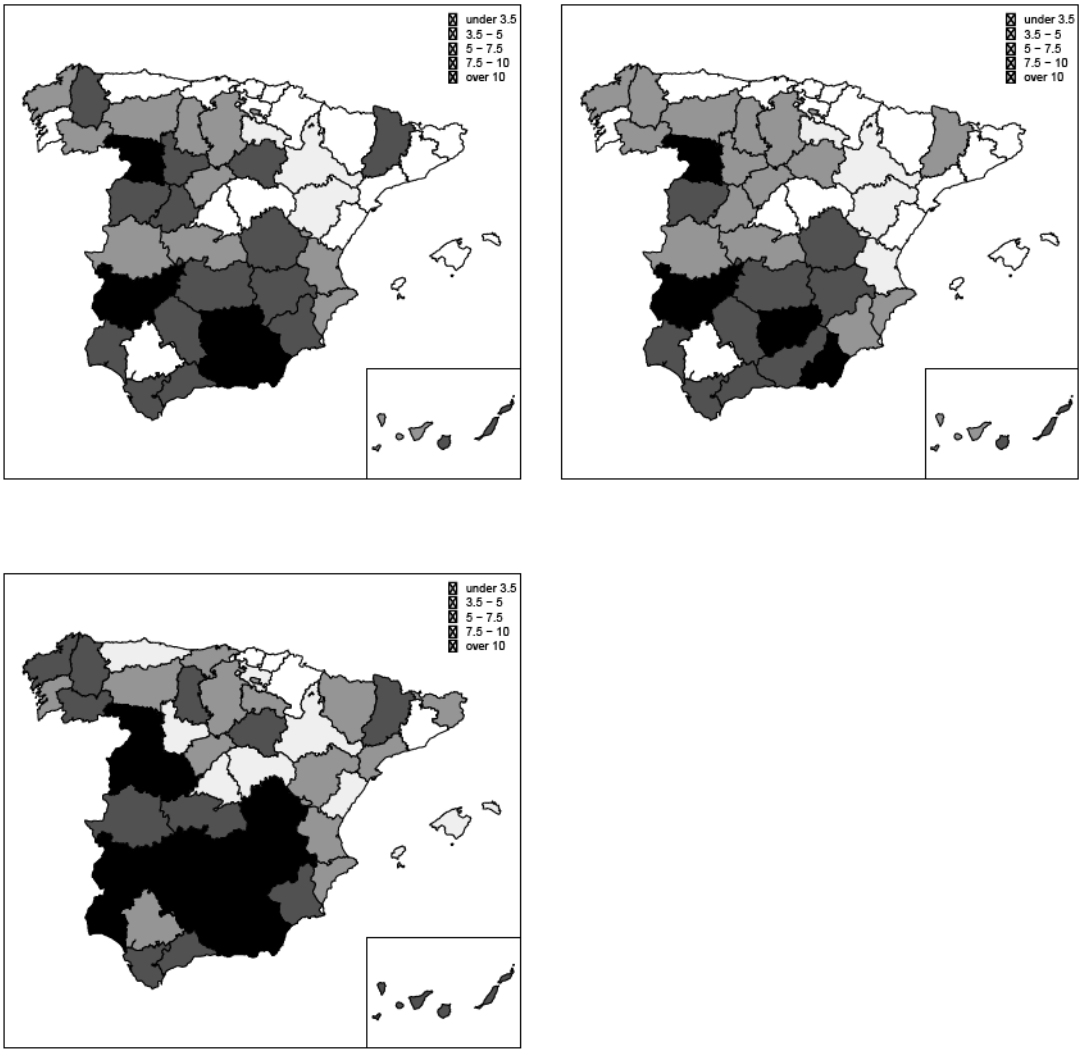


Figure 2.11: Cartograms of estimated percent poverty gaps, $\hat{F}_{1i}$, in Spanish provinces for men obtained with direct (top left), FH (top right) and HB (bottom left) methods.



Figures 2.10 and 2.11 show the analogous cartograms for men. All the methods coincide in that the largest poverty incidences are for the provinces of Zamora, Badajoz and Almería. Again, the FH estimator tends to smooth the direct estimates. For the poverty gap, the three estimates coincide in the provinces of Zamora, Badajoz, Almería and Jaén as the ones with the largest values. As in poverty incidence, we can see that the FH estimator tends to smooth the direct estimates.

Before concluding this chapter, we have to take into account that, in this example, we do not really know if there is a problem of informative sampling. In fact, non-response can be seen as a problem of informative selection if the probability of responding is related with the target variable, in this case income. According to the documentation provided by Spanish National Institute of Statistics, in 2011, the SILC survey presented a rate of household non-response of about 35%. The individual non-response rate was of about 35.5% for the same year. If the non-respondents do not follow the same model as the respondents, then HB and EB estimates might be biased. Since sampling weights are available and are adjusted for non-response, the problem may be analyzed by studying if these sampling weights are related with the incomes. In this case, we would need to apply new model-based estimation methods that incorporate the sampling weights. We propose a new method of this kind in Chapter 3 and apply the new method to Mexican data in Section 3.7.

Chapter 3

Small area estimation of general parameters under informative sampling

For the estimation of general non-linear parameters in small domains or areas, Molina and Rao (2010) introduced the empirical best (EB) method based on the unit level nested error model of Battese et al. (1988). Non-linear parameters of great interest are poverty or inequality indicators, and reliable estimates of this kind of indicators at regional level can be used to construct maps showing the regional distribution of poverty or inequality in a certain population or country. The World Bank has been producing poverty maps for many countries all over the world using traditionally the method of Elbers et al. (2003), called here ELL method. Under the same model assumptions, EB method for poverty mapping outperforms ELL method when the area effects are significant, see Molina and Rao (2010). Both methods assume that the model for the sampled units is exactly the same as the model considered for the population; in other words, the sample selection mechanism is not affecting the distribution of the outcomes (non-informative selection). In the case of informative selection, using the sample without appropriate weighting to obtain EB estimators of poverty indicators leads to biased estimators.

In the literature, we can find few approaches to handle informative selection in small area estimation. The approach of Pfeffermann and Sverchkov (2007) is to calculate the sample likelihood as the usual likelihood conditional on the selected sample, where the inclusion probabilities are modeled in terms of the observed outcomes and covariates. The sample likelihood is used to obtain maximun likelihood estimators of the model parameters that are corrected by the informativeness of the sample selection and these estimates are then used to estimate domain means. This procedure has not been

extended for estimation of non-linear parameters. The approach of Verret et al. (2015) is to model the outcomes in terms of the sampling weights or inclusion probabilities and covariates, that is, to augment the assumed population model for the outcomes by including the weights or inclusion probabilities as an additional covariate. Applying this augmenting model approach for non-linear parameters would require to have the inclusion probabilities or sampling weights not only for the sample units, but for the non-sample units as well.

To handle complex sampling designs (not necessarily informative), we can find several approaches that incorporate the sampling weights to obtain design consistent estimators when estimating linear domain parameters. Prasad and Rao (1999) and You and Rao (2002) propose a weighted version of the empirical best linear unbiased predictor (EBLUP) using survey weights, called pseudo EBLUP. Lehtonen and Veijanen (1999) propose the multilevel-model assisted generalized regression estimator (MGREG), which is a GREG estimator assisted by a multilevel model. Lehtonen et al. (2003) propose a generalization of the MGREG, which includes a GREG assisted by a logistic mixed model. Fabrizi et al. (2014) propose to use sampling weights in the context of small area estimation with M-Quantiles.

In this chapter, we introduce a new procedure that reduces the bias due to an informative selection mechanism based on combining the ideas of conditioning on the sample of the EB method with the correct weighting of design-based estimators. Instead of conditioning on the sample mean of the target area as EB method does, we propose to condition on the weighted sample mean using as weights the inverses of the inclusion probabilities. This leads to a weighted EB approach called here pseudo EB.

This chapter is organized as follows. Section 3.1 introduces the assumed population model. Section 3.2 defines informative/non-informative selection mechanisms. EB method is reviewed in Section 3.3 and our proposal is described in Section 3.4. A bootstrap procedure for mean squared error estimation is included in Section 3.5. Results of simulation experiments carried out under both informative and non-informative selection are described in Section 3.6. Finally, Section 3.7 applies the proposed method to poverty mapping in the municipalities from the State of Mexico and compares the resulting estimates with the unweighted EB estimates.

3.1 Population model

In this chapter, we wish to estimate a certain characteristic in each of m domains or areas U_i, $i = 1, \ldots, m$, into which our finite population U is partitioned. The size of domain U_i is N_i, $i = 1, \ldots, m$, where $N = \sum_{i=1}^{m} N_i$ is the total population size. We denote by Y_{ij}

the measurement of the study variable for j-th unit within i-th domain.

We assume that the population measurements Y_{ij} follow the nested error model introduced by Battese et al. (1988),

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij},\ e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2),\ j = 1, \ldots, N_i, \tag{3.1}$$

$$v_i \overset{iid}{\sim} N(0, \sigma_v^2),\ i = 1, \ldots, m, \tag{3.2}$$

where $\mathbf{x}_{ij}$ is a $p \times 1$ vector of auxiliary variables, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, v_i is the effect (unexplained heterogeneity) of domain i and e_{ij} is the individual regression error, where domain effects and errors are all mutually independent. Let us write the model in matrix notation by defining the domain vectors and matrices

$$\mathbf{y}_i = (Y_{i1}, \ldots, Y_{iN_i})',\ \mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i})',\ \mathbf{e}_i = (e_{i1}, \ldots, e_{iN_i})',\ i = 1, \ldots, m.$$

Then, model (3.1)-(3.2) becomes

$$\mathbf{y}_i \overset{ind}{\sim} N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i),\ \mathbf{V}_i = \sigma_v^2 \mathbf{1}_{N_i}\mathbf{1}'_{N_i} + \sigma_e^2 \mathbf{I}_{N_i},\ i = 1, \ldots, m, \tag{3.3}$$

where $\mathbf{1}_k$ denotes a vector of ones of size k and $\mathbf{I}_k$ is the $k \times k$ identity matrix. Additionally, we denote by $\mathbf{y} = (\mathbf{y}'_1, \ldots, \mathbf{y}'_m)'$ the population vector of measurements, $\mathbf{X} = (\mathbf{X}'_1, \ldots, \mathbf{X}'_m)'$ is the population design matrix and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_v^2, \sigma_e^2)'$ is the vector of unknown model parameters.

We wish to estimate possibly non-linear domain parameters that are additive in the population units, in the sense that they can be expressed as

$$H_i = \frac{1}{N_i} \sum_{j=1}^{N_i} h(Y_{ij}), \quad i = 1, \ldots, m, \tag{3.4}$$

where $h(\cdot)$ is a real function. For the special case $h(y) = y$, we obtain the mean of domain i, that is, $H_i = \bar{Y}_i$.

3.2 Sample selection mechanism

The target domain parameters H_i, $i = 1, \ldots, m$, are estimated based on a sample s drawn from the population U using a given selection mechanism or sampling design. The sample s is composed of subsamples s_i, drawn independently from each domain U_i, $i = 1, \ldots, m$. Let n_i be the sample size of domain i, $i = 1, \ldots, m$. The total sample size is then $n = \sum_{i=1}^{m} n_i$. We denote by $r_i = U_i - s_i$ the set of out-of-sample units from domain i, of size $N_i - n_i$, $i = 1, \ldots, m$. In this chapter, we assume that the population

matrix $\mathbf{X}$ of auxiliary variables is available from a census or a register. Then, all the probability distributions involved in this paper are conditional on $\mathbf{X}$ but we will omit this dependence in the notation for simplicity.

Traditional model-based inference assumes that the selection mechanism is non-informative. This means that the probability of the sample is not related with the outcome values. More formally, let $P(s|\mathbf{y})$ be the probability of sample s according to the selected sampling mechanism given $\mathbf{y}$ (and $\mathbf{X}$). We say that the sampling design is non-informative for inference on characteristics of $\mathbf{y}$ when

$$P(s|\mathbf{y}) = P(s), \quad \forall \mathbf{y} \in I\!R^N, \forall s.$$

Equivalently, using Bayes Theorem, the sampling is non-informative for inference on characteristics of $\mathbf{y}$ when

$$f(\mathbf{y}|s) = f(\mathbf{y}), \quad \forall\, \mathbf{y} \in I\!R^N, \forall\, s.$$

Otherwise, we say that the sampling design is informative for inference about $\mathbf{y}$. Let $\mathbf{y}_s$ be the sub-vector of $\mathbf{y}$ corresponding to the sample units. Under non-informative sampling, $f(\mathbf{y}_s|s) = f(\mathbf{y}_s)$ and then inference based on the usual likelihood $f(\mathbf{y}_s)$ is valid. This means that the selection process does not affect the distribution of the outcomes for selected units.

There are weaker conditions for ignorability of the sampling design based on auxiliary information, see e.g. Section 2 of Pfeffermann and Sverchkov (2009) and the references therein. Here we consider that the available auxiliary information is not enough for ensuring that the design is ignorable.

3.3 EB method

This method assumes that the sampling design is non-informative for inference about $\mathbf{y}$. Then, the outcomes corresponding to sampled units, Y_{ij}, $j \in s_i$, preserve the same distribution as the outcomes for out-of-sample units, given by (3.1)-(3.2) under the considered nested error model. Let us decompose the domain vector $\mathbf{y}_i$ into subvectors corresponding to sample and out-of-sample elements as $\mathbf{y}_i = (\mathbf{y}'_{is}, \mathbf{y}'_{ir})'$, where the subscript s denotes the sample units and r the out-of-sample units. The sample data is then $\mathbf{y}_s = (\mathbf{y}'_{1s}, \ldots, \mathbf{y}'_{ms})'$. For a general domain parameter $H_i = h(\mathbf{y}_i)$, the best predictor is defined as the function of the sample observations $\mathbf{y}_s$ that minimizes the

mean squared error (MSE) and is given by

$$\tilde{H}_i^B(\boldsymbol{\theta}) = E_{\mathbf{y}_{ir}}(H_i|\mathbf{y}_{is};\boldsymbol{\theta}),$$

where the expectation is taken with respect to the distribution of $\mathbf{y}_{ir}|\mathbf{y}_{is}$, which depends on the true value of $\boldsymbol{\theta}$. For a domain parameter H_i that is additive as in (3.4), the best predictor is reduced to

$$\tilde{H}_i^B(\boldsymbol{\theta}) = \frac{1}{N_i}\left[\sum_{j\in s_i} h(Y_{ij}) + \sum_{j\in r_i} \tilde{H}_{ij}^B(\boldsymbol{\theta})\right], \tag{3.5}$$

where $\tilde{H}_{ij}^B(\boldsymbol{\theta}) = E[h(Y_{ij})|\mathbf{y}_{is};\boldsymbol{\theta}]$ is also the best predictor of $H_{ij} = h(Y_{ij})$ for out-of-sample unit $j \in r_i$. The best predictor $\tilde{H}_i^B(\boldsymbol{\theta})$ is exactly model unbiased for H_i regardless of the complexity of the function $h(\cdot)$. However, it cannot be calculated in practice since model parameters $\boldsymbol{\theta}$ are typically unknown. An empirical best predictor (EB) of H_i, denoted as $\hat{H}_i^{EB}$, is then obtained by replacing $\boldsymbol{\theta}$ in $\tilde{H}_i^B(\boldsymbol{\theta})$ by a consistent estimator $\hat{\boldsymbol{\theta}}$, that is, $\hat{H}_i^{EB} = \tilde{H}_i^B(\hat{\boldsymbol{\theta}})$. The EB predictor is not exactly unbiased, but the bias arising from the estimation of $\boldsymbol{\theta}$ is typically negligible when the overall sample size n is large. For $h(\cdot)$ linear and under normality of $\mathbf{y}$, the EB predictor of H_i equals the empirical best linear unbiased predictor (EBLUP) of H_i.

Given the nested error model specified in (3.1)-(3.2) and assuming non-informative selection, the out-of-sample vectors $\mathbf{y}_{ir}$ given the sample data vectors $\mathbf{y}_{is}$ are independent and follow exactly the same distribution as $\mathbf{y}_{ir}|\bar{y}_{is}$, where $\bar{y}_{is}$ is the unweighted sample mean for area i. Thus, the best predictor of $H_{ij} = h(Y_{ij})$ is $\tilde{H}_{ij}^B(\boldsymbol{\theta}) = E[h(Y_{ij})|\bar{y}_{is};\boldsymbol{\theta}]$. For an out-of-sample observation Y_{ij}, $j \in r_i$, we have

$$Y_{ij}|\bar{y}_{is} \sim N(\mu_{ij|s}, \sigma^2_{ij|s}), \quad j \in r_i, \tag{3.6}$$

$$\mu_{ij|s} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_{is}(\bar{y}_{is} - \bar{\mathbf{x}}'_{is}\boldsymbol{\beta}), \quad \sigma^2_{ij|s} = \sigma_v^2(1-\gamma_{is}) + \sigma_e^2, \tag{3.7}$$

for $\bar{\mathbf{x}}_{is} = n_i^{-1}\sum_{j\in s_i}\mathbf{x}_{ij}$ and $\gamma_{is} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2/n_i)$.

Foster et al. (1984) introduced the family of FGT poverty indicators, which contain several widely-used poverty measures and which are additive in the sense described above. In particular, the poverty maps released by World Bank are traditionally based on members of this family. Let E_{ij} be a welfare measure for individual j in area i and z be the poverty line. The family of FGT poverty indicators for domain i is given by

$$F_{\alpha i} = \frac{1}{N_i}\sum_{j=1}^{N_i} F_{\alpha ij}, \quad F_{\alpha ij} = \left(\frac{z - E_{ij}}{z}\right)^{\alpha} I(E_{ij} < z), \quad j = 1,\ldots,N_i, \alpha \geq 0, \tag{3.8}$$

where $I(E_{ij} < z) = 1$ if $E_{ij} < z$, and $I(E_{ij} < z) = 0$ otherwise. For $\alpha = 0$, we obtain the poverty incidence, measuring the frequency of poverty. For $\alpha = 1$, we get the poverty gap, measuring the poverty depth. Both indicators together give a good description of poverty.

Consider that the model (3.1)-(3.2) holds for $Y_{ij} = \log(E_{ij} + c)$, where $c \geq 0$ is a constant. Then, we can express $F_{\alpha ij}$ in terms of the response variable Y_{ij} as

$$F_{\alpha ij} = \left[\frac{z - \exp(Y_{ij}) + c}{z}\right]^{\alpha} I[\exp(Y_{ij}) - c < z] =: h_{\alpha}(Y_{ij}),$$

which shows that $F_{\alpha i} = N_i^{-1} \sum_{j=1}^{N_i} h_{\alpha}(Y_{ij})$ is an additive parameter in the sense of (3.4). According to (3.5), the best predictor of $H_i = F_{\alpha i}$ is then given by

$$\tilde{F}_{\alpha i}^{B}(\boldsymbol{\theta}) = \frac{1}{N_i}\left(\sum_{j \in s_i} F_{\alpha ij} + \sum_{j \in r_i} \tilde{F}_{\alpha ij}^{B}(\boldsymbol{\theta})\right), \tag{3.9}$$

where $\tilde{F}_{\alpha ij}^{B}(\boldsymbol{\theta}) = E[h_{\alpha}(Y_{ij})|\bar{y}_{is}; \boldsymbol{\theta}]$ is the best predictor of $F_{\alpha ij} = h_{\alpha}(Y_{ij})$. For $\alpha = 0, 1$, the best predictor $\tilde{F}_{\alpha ij}^{B}(\boldsymbol{\theta})$ can be calculated analytically. Let us define $\alpha_{ij} = [\log(z+c) - \mu_{ij|s}]/\sigma_{ij|s}$, for $\mu_{ij|s}$ and $\sigma_{ij|s}^2$ given in (3.6)-(3.7). Then, the best predictors of F_{0ij} and F_{1ij} are respectively given by

$$\tilde{F}_{0ij}^{B}(\boldsymbol{\theta}) = \Phi(\alpha_{ij}), \tag{3.10}$$

$$\tilde{F}_{1ij}^{B}(\boldsymbol{\theta}) = \Phi(\alpha_{ij})\left\{1 - \frac{1}{z}\left[\exp\left(\mu_{ij|s} + \frac{\sigma_{ij|s}^2}{2}\right)\frac{\Phi(\alpha_{ij} - \sigma_{ij|s})}{\Phi(\alpha_{ij})} - c\right]\right\}, \tag{3.11}$$

where $\Phi(\cdot)$ is the c.d.f. of a standard Normal random variable.

For additive area parameters $H_i = N_i^{-1} \sum_{j=1}^{N_i} h(Y_{ij})$ with more complex $h(\cdot)$, analytical expressions for the expectation $E[h(Y_{ij})|\bar{y}_{is}; \boldsymbol{\theta}]$ defining the best predictor may not be available. In any case, the EB predictor $\hat{H}_{ij}^{EB} = E[h(Y_{ij})|\bar{y}_{is}; \hat{\boldsymbol{\theta}}]$ of a general $H_{ij} = h(Y_{ij})$ can be approximated by Monte Carlo, similarly as in Molina and Rao (2010). This is done by simulating L replicates $\{Y_{ij}^{(\ell)}; \ell = 1, \ldots, L\}$ of Y_{ij}, $j \in r_i$, from the estimated conditional distribution of $Y_{ij}|\bar{y}_{is}$ given in (3.7), calculating the corresponding $h(Y_{ij}^{(\ell)})$ for each ℓ and then averaging over the L replicates as $\hat{H}_{ij}^{EB} = L^{-1} \sum_{\ell=1}^{L} h(Y_{ij}^{(\ell)})$.

When the sample units cannot be identified in the census of auxiliary variables, EB estimators given by (3.9) with $\boldsymbol{\theta}$ replaced by a consistent estimator $\hat{\boldsymbol{\theta}}$, cannot be calculated. A variation of the EB estimator, called Census EB estimator, is obtained by predicting the sample values H_{ij}, $j \in s_i$, pretending as if they were out of the sample to

obtain $\hat{H}_{ij}^{EB}$ for them, and then taking

$$\hat{H}_i^{CEB} = \frac{1}{N_i}\sum_{j=1}^{N_i} \hat{H}_{ij}^{EB}, \tag{3.12}$$

see Guadarrama et al. (2016). Typically the sampling fraction n_i/N_i is very small. In that case, the census EB estimator of H_i is approximately equal to the EB estimator.

In the Appendix B, we show that under simple random sampling within area i and for known model parameters $\boldsymbol{\theta}$, the Census EB estimator $\hat{F}_{\alpha i}^{CEB}$ of the poverty indicator $F_{\alpha i}$ for $\alpha = 0, 1$, is consistent as $n_i \to \infty$ and $N_i \to \infty$ under the joint distribution of the sampling design and model (3.1) for $Y_{ij}|v_i$, without any assumption on the distribution of the domain effects v_i.

3.4 Pseudo EB method

As stated above, under the nested error model (3.1)-(3.2), $\mathbf{y}_{ir}|\bar{y}_{is}$ follows exactly the same distribution as $\mathbf{y}_{ir}|\mathbf{y}_{is}$ and the best predictor of $H_{ij} = h(Y_{ij})$, $j \in r_i$, can be expressed as $\tilde{H}_{ij}^{B} = E[h(Y_{ij})|\bar{y}_{is}]$. When the sample selection mechanism is informative, to avoid a bias due to a non-representative sample, the estimation procedure should incorporate the sampling weights. Let w_{ij} be the sampling weight of j-th unit within i-th domain and $w_{i\cdot} = \sum_{j\in s_i} w_{ij}$. We consider the same conditioning idea of the EB estimator, but now we condition on the weighted sample mean $\bar{y}_{iw} = w_{i\cdot}^{-1}\sum_{j\in s_i} w_{ij}y_{ij}$ instead of the unweighted sample mean $\bar{y}_{is}$. Thus, we define the pseudo best (PB) estimator of $H_{ij} = h(Y_{ij})$ as

$$\tilde{H}_{ij}^{PB}(\boldsymbol{\theta}) = E[h(Y_{ij})|\bar{y}_{iw};\boldsymbol{\theta}]. \tag{3.13}$$

The PB estimator of the additive area parameter H_i is then

$$\tilde{H}_i^{PB}(\boldsymbol{\theta}) = \frac{1}{N_i}\left[\sum_{j\in s_i} h(Y_{ij}) + \sum_{j\in r_i} \tilde{H}_{ij}^{PB}(\boldsymbol{\theta})\right]. \tag{3.14}$$

Jiang and Lahiri (2006) used a similar approach in the special case of area means under the nested error model and also in the case of a binary response variable and a logit linking model. Their method is applicable only for area level covariates in the unit level models. For example, when using the area mean vector $\bar{\mathbf{X}}_i = N_i^{-1}\sum_{i=1}^{N_i}\mathbf{x}_{ij}$ as area level covariates in the unit level model.

Similarly as in EB method, the PB estimator (3.14) depends on the true values of the

model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_v^2, \sigma_e^2)'$, which need to be estimated. We define the pseudo EB (PEB) predictor as the PB predictor with $\boldsymbol{\theta}$ replaced by a consistent estimator. The approach of Pfeffermann and Sverchkov (2007) based on the sample likelihood can be used to find correct maximum likelihood (ML) estimates of the regression parameter $\boldsymbol{\beta}$ and of the variances σ_v^2 and σ_e^2. Alternatively, $\boldsymbol{\beta}$ can be estimated using the weighted method of moments used in You and Rao (2002) and using ML (or REML) estimators of σ_v^2 and σ_e^2.

For an out-of-sample variable Y_{ij}, $j \in r_i$, under the nested error population model (3.1)-(3.2), we have

$$Y_{ij}|\bar{y}_{iw} \overset{ind.}{\sim} N(\mu^w_{ij|s}, \sigma^{2w}_{ir|s}), \tag{3.15}$$

$$\mu^w_{ij|s} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_{iw}(\bar{y}_{iw} - \bar{\mathbf{x}}'_{iw}\boldsymbol{\beta}), \quad \sigma^{2w}_{ij|s} = \sigma_v^2(1-\gamma_{iw}) + \sigma_e^2, \tag{3.16}$$

where $\bar{\mathbf{x}}_{iw} = w_{i\cdot}^{-1}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}$ and $\gamma_{iw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2\delta_i^2)$, for $\delta_i^2 = w_{i\cdot}^{-2}\sum_{j\in s_i} w_{ij}^2$. Observe that the mean $\mu^w_{ij|s}$ is obtained from $\mu_{ij|s}$ given in (3.7) by replacing the unweighted best predictor $\tilde{v}_{is} = \gamma_{is}(\bar{y}_{is} - \bar{\mathbf{x}}'_{is}\boldsymbol{\beta})$ of the domain effect v_i by its weighted version, given by $\tilde{v}_{iw} = \gamma_{iw}(\bar{y}_{iw} - \bar{\mathbf{x}}'_{iw}\boldsymbol{\beta})$. Even if the conditional distribution (3.15)–(3.16) is obtained assuming that the sample units satisfy the same population model (3.1)–(3.2) (i.e. non-informative sampling), we will see that conditioning on the weighted sample mean $\bar{y}_{iw}$ protects against informative sampling.

For the FGT poverty indicators of order $\alpha = 0, 1$, the PB are given by (3.10) and (3.11) with $\mu_{ij|s}$ and $\sigma^2_{ij|s}$ replaced by the weighted versions $\mu^w_{ij|s}$ and $\sigma^{2w}_{ij|s}$. For more complex additive parameters, such as the FGT indicators for $\alpha > 1$, we can apply a Monte Carlo procedure to approximate the PEB predictor of $H_{ij} = h(Y_{ij})$ similarly as done for the EB predictor. We generate L replicates $\{Y_{ij}^{(\ell)}; \ell = 1, \ldots, L\}$ of Y_{ij}, $j \in r_i$, from the estimated conditional distribution of $Y_{ij}|\bar{y}_{iw}$ given in (3.15)–(3.16), calculate $h(Y_{ij}^{(\ell)})$ for each ℓ and then average over the L replicates as $\hat{H}_{ij}^{PEB} = L^{-1}\sum_{\ell=1}^{L} h(Y_{ij}^{(\ell)})$.

Similarly as in the Census EB estimator given in (3.12), we define the Census PEB estimator as

$$\hat{H}_i^{CPEB} = \frac{1}{N_i}\sum_{j=1}^{N_i} \hat{H}_{ij}^{PEB}. \tag{3.17}$$

Note that the Census PEB estimator (3.17) is obtained by predicting also the sample values $H_{ij} = h(Y_{ij})$ as if they were out of sample. Under general sampling designs, in Appendix B we show that, for known $\boldsymbol{\theta}$, the Census PEB estimator $\hat{F}_{\alpha i}^{CPEB}$ of the poverty indicator $F_{\alpha i}$, for $\alpha = 0, 1$, is consistent as $n_i \to \infty$ and $N_i \to \infty$, under the joint distribution of the sampling design and the considered model for $Y_{ij}|v_i$ given in (3.1), without any assumption on the distribution of v_i.

For the special case of a domain mean $H_i = \bar{Y}_i$, if $\boldsymbol{\beta}$ is estimated by the weighted regression estimator $\hat{\boldsymbol{\beta}}_w$ given in You and Rao (2002), the Census PEB estimator of $H_i = \bar{Y}_i$ equals the pseudo EBLUP of You and Rao (2002). Similarly, the PEB estimator obtained from (3.14) tends to the pseudo EBLUP as the domain sampling fraction $f_i = n_i/N_i$ becomes small. Thus, for a domain mean $\bar{Y}_i$, the Census PEB estimator (and PEB for small domain sampling fraction) preserves the good properties of the pseudo EBLUP: a) design consistency as n_i becomes large, and b) automatic benchmarking to the survey regression estimator of the overall population total, provided the sampling weights are calibrated to agree with the known population total $w_{i\cdot} = N_i$. Stefan et al. (2005) and Verret et al. (2015) showed that the pseudo EBLUP of the area mean $\bar{Y}_i$ performs well under informative sampling in terms of bias and mean squared error (MSE) under the model.

3.5 Parametric bootstrap MSE estimator

Even though the PEB estimators proposed in Section 3.4 incorporate the sampling weights, they are essentially model-based. Thus, here we propose estimators of the MSE of PEB estimators under the model. We consider a similar bootstrap procedure as in Molina and Rao (2010), based on the parametric bootstrap method for finite populations introduced by González-Manteiga et al. (2008). The parametric bootstrap estimator of the model MSE of $\hat{H}_i^{PEB}$ is obtained as follows: i) Fit the model (3.1)-(3.2) to the sample data $(\mathbf{y}_s, \mathbf{X}_s)$ and obtain estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ of $\boldsymbol{\beta}$, σ_v^2 and σ_e^2 respectively. ii) For $b = 1, \ldots, B$, with B large, generate $v_i^{*(b)} \sim N(0, \hat{\sigma}_v^2)$ and $e_{ij}^{*(b)} \sim N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$, independently. iii) Construct B iid bootstrap population vectors $\mathbf{y}^{*(b)}$, $b = 1, \ldots, B$, with elements $Y_{ij}^{*(b)}$ generated as

$$Y_{ij}^{*(b)} = \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} + v_i^{*(b)} + e_{ij}^{*(b)}, \quad j = 1, \ldots, N_i,\ i = 1, \ldots, m.$$

From each bootstrap population b, calculate the true value of the domain parameter $H_i^{*(b)} = N_i^{-1} \sum_{j=1}^{N_i} h(Y_{ij}^{*(b)})$, $b = 1, \ldots, B$. iv) From each bootstrap population b, take the sample with the same indices as the initial sample s and, using the sample elements $\mathbf{y}_s^{*(b)}$ of $\mathbf{y}^{*(b)}$ and the known population vectors $\mathbf{x}_{ij}$, $j \in U_i$, calculate the bootstrap pseudo EB predictors of H_i, denoted $\hat{H}_i^{PEB*(b)}$, $b = 1, \ldots, B$. v) A bootstrap estimator of the model MSE of the PEB estimator, $\text{MSE}_m(\hat{H}_i^{PEB})$, is then

$$\text{mse}_m(\hat{H}_i^{PEB}) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{H}_i^{PEB*(b)} - H_i^{*(b)} \right)^2. \tag{3.18}$$

3.6 Simulation experiments

We carried out simulation experiments to analyze the performance of the PEB estimators $\hat{F}_{\alpha i}^{PEB}$ of poverty incidences and gaps, $F_{\alpha i}$, $\alpha = 0, 1$, compared to EB estimators $\hat{F}_{\alpha i}^{EB}$. We also compare with two types of direct estimators. Since the considered poverty indicators are population means $F_{\alpha i} = N_i^{-1} \sum_{j=1}^{N_i} F_{\alpha ij}$, we can calculate the usual (unweighted) sample mean (SM) of $F_{\alpha ij}$, for $j \in s_i$, as well as the weighted sample mean (WSM) of $F_{\alpha ij}$, for $j \in s_i$, that is,

$$\bar{F}_{\alpha i} = \frac{1}{n_i} \sum_{j \in s_i} F_{\alpha ij}, \quad \bar{F}_{\alpha i,w} = \frac{1}{w_{i\cdot}} \sum_{j \in s_i} w_{ij} F_{\alpha ij}. \tag{3.19}$$

We wish to analyze the performance of our model-based estimators under general selection mechanisms including informative ones. For this reason, our simulation experiments will be under a model-design setup, that is, with respect to the joint distribution of the population vector $\mathbf{y}$ and the sample s, $(\mathbf{y}, s)$. In each Monte Carlo (MC) simulation, a population vector $\mathbf{y}$ is generated according to model (3.1)-(3.2) and a sample s is drawn according to a given selection mechanism. In Section 3.6.1, we give the results of a simulation experiment where the sample is drawn by a complex but non-informative mechanism. Section 3.6.2 shows the results of a simulation study where the sample is drawn by an informative selection mechanism.

3.6.1 Simulation study with non-informative selection

We consider the same simulation setup as in Chapter 2, where the population contains $N = 20,000$ units distributed into $m = 80$ domains, with $N_i = 250$ units in each domain $i = 1, \ldots, m$. We consider two dummy auxiliary variables, $x_q \in \{0, 1\}$, $q = 1, 2$, whose values are generated as $x_{q,ij} \sim \text{Bern}(p_{qi})$, $q = 1, 2$, with success probabilities given by $p_{1i} = 0.3 + 0.5\, i/m$ and $p_{2i} = 0.2$, $i = 1, \ldots, m$. The values of the auxiliary variables $x_{q,ij}$ are kept fixed across simulations. The vector of true regression coefficients is taken as $\boldsymbol{\beta} = (3, 0.03, -0.04)'$ and the domain effects variance and error variance are respectively $\sigma_v^2 = 0.15^2$ and $\sigma_e^2 = 0.5^2$.

In each MC simulation out of $K = 1,000$, we construct a population vector $\mathbf{y}^{(k)}$, whose elements $Y_{ij}^{(k)}$ are generated from the nested error model (3.1)-(3.2). Using the population vector $\mathbf{y}^{(k)}$, we calculate the true values of the domain parameters $F_{\alpha i}^{(k)}$, $i = 1, \ldots, m$. We take the poverty line as $z = 12$, which is approximately 0.6 times the median of a population of incomes $\{E_{ij}; j = 1, \ldots, N_i,\ i = 1, \ldots, m\}$, where $E_{ij} = \exp(Y_{ij})$ with Y_{ij} from nested error model generated as mentioned above. For each MC population $k = 1, \ldots, K$, we draw a sample $s^{(k)}$. We use independent Poisson

sampling within each domain i, with inclusion probability for individual j in the sample from domain i taken as $\pi_{ij} \sim \text{Beta}(\alpha_1, \alpha_2)$. We set $\alpha_1 = 2.5$ and select α_2 to achieve a specified expected domain sample size, $\bar{n}_i = K^{-1}\sum_{k=1}^{K} n_i^{(k)}$, where $n_i^{(k)}$ is the realized sample size in domain i in the k-th MC replicate. We consider three expected domain sample sizes: $\bar{n}_i = 25, 50, 75$. To achieve approximately those domain sample sizes, we take $\alpha_2 = 25$, $\alpha_2 = 10$ and $\alpha_2 = 5.5$ respectively. We consider that the proposed sample sizes are small enough since we are estimating small proportions.

With the sample data from the k-th Monte Carlo population $\mathbf{y}_s^{(k)}$, we compute two direct estimators of $F_{\alpha i}^{(k)}$, namely the SM and also the WSM as in (3.19), using as weights $w_{ij} = \pi_{ij}^{-1}$. We also compute EB and pseudo EB estimates of $F_{\alpha i}^{(k)}$, for $\alpha = 0, 1$ and $i = 1, \ldots, m$, using the population values of the auxiliary variables. For the EB estimator, we computed $\hat{\sigma}_v^2$, $\hat{\sigma}_e^2$ and $\hat{\boldsymbol{\beta}}$ by the REML method. For the pseudo EB estimator, we used the weighted estimator $\hat{\boldsymbol{\beta}}_w$ given in You and Rao (2002) and the REML estimators of σ_v^2 and σ_e^2. Let $\hat{F}_{\alpha i}^{(k)}$ be one of the mentioned estimates (SM, WSM, EB or pseudo EB) in MC replicate k. We evaluate the performance of estimators in terms of relative bias (RB) and relative root MSE (RRMSE), under the model and the design, approximated empirically as

$$\text{RB}_{m,\pi}(\hat{F}_{\alpha i}) = \frac{K^{-1}\sum_{k=1}^{K}\left(\hat{F}_{\alpha i}^{(k)} - F_{\alpha i}^{(k)}\right)}{K^{-1}\sum_{k=1}^{K} F_{\alpha i}^{(k)}}, \quad \text{RRMSE}_{m,\pi}(\hat{F}_{\alpha i}) = \frac{\sqrt{K^{-1}\sum_{k=1}^{K}\left(\hat{F}_{\alpha i}^{(k)} - F_{\alpha i}^{(k)}\right)^2}}{K^{-1}\sum_{k=1}^{K} F_{\alpha i}^{(k)}}.$$

Averages across domains of absolute RB and of RRMSE are also calculated as

$$\overline{\text{ARB}}_\alpha = m^{-1}\sum_{i=1}^{m} |\text{RB}_{m,\pi}(\hat{F}_{\alpha i})|, \quad \overline{\text{RRMSE}}_\alpha = m^{-1}\sum_{i=1}^{m} \text{RRMSE}_{m,\pi}(\hat{F}_{\alpha i}).$$

Figures 3.1, 3.2 and 3.3 display, respectively for approximate expected domain sample sizes $\bar{n}_i = 25, 50$ and 75, percent RB (left) and RRMSE (right) of the estimators of the poverty gap, F_{1i}, for each domain $i = 1, \ldots, m$ (x-axis). In these figures, all the estimators display a small RB for the three expected sample sizes, although the WSM appears to be more unstable across domains than the other ones. This estimator also performs the worst in terms of RRMSE, followed by the unweighted SM. Thus, model-based estimators (EB and pseudo EB) appear to be significantly more efficient than the two types of direct estimators (SM and WSM) for all the domains. In this simulation experiment with non-informative sampling, weighted estimators (WSM and pseudo EB) loose efficiency with respect to the respective unweighted ones, but the

efficiency loss of the pseudo EB turns out to be much smaller than the loss of the WSM with respect to the SM. As expected, the gain in efficiency of the model-based estimators compared to the direct estimators decreases as the expected sample size increases, with SMs becoming close to model-based estimators for the largest expected domain sample size $\bar{n}_i$ (Figure 3.3). Conclusions for the poverty incidence, F_{0i}, are similar and hence figures are not shown.

Table 3.1 displays averages of absolute RB and RRMSE across domains for the considered expected domain sample sizes. This table shows an $\overline{\text{ARB}}$ smaller than 2% for all the considered estimators and sample sizes. EB and pseudo EB estimators have considerably smaller $\overline{\text{RRMSE}}$ than direct estimators for small $\bar{n}_i$ and preserve smaller $\overline{\text{RRMSE}}$ even for the largest value of $\bar{n}_i$. Since the sample selection mechanism is in this case non-informative, the $\overline{\text{RRMSE}}$ of pseudo EB estimator turns out to be between 3% and 4% larger than that of EB estimator. This suggests that EB estimators work well under unequal probability sampling as long as the inclusion probabilities do not depend on the outcomes. Nevertheless, in this case pseudo EB estimator does not lose too much.

	$\bar{n}_i = 25$				$\bar{n}_i = 50$				$\bar{n}_i = 75$			
	$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$		$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$		$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$	
Method	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}
SM	1.34	1.65	46.27	58.69	0.69	0.87	29.03	36.85	0.54	0.66	21.41	27.93
WSM	1.65	1.94	56.46	71.59	0.83	1.12	36.26	45.95	0.68	0.82	26.98	34.34
EB	0.74	0.89	28.21	35.60	0.46	0.60	20.99	26.73	0.40	0.47	17.58	22.29
PEB	0.88	1.04	31.25	39.29	0.54	0.72	24.13	30.43	0.49	0.61	20.07	25.39

Table 3.1: Averages across domains of percent absolute RB and RRMSE for SM, WSM, EB and pseudo EB estimators of poverty incidence, F_{0i}, and poverty gap, F_{1i}, under non-informative selection with $\bar{n}_i = 25, 50, 75$.

3.6.2 Simulation study with informative selection

A simulation experiment was carried out with the same population structure and the same model that generates the population values as in Section 3.6.1. However, in this experiment, for each MC replicate, we draw the sample using an informative selection mechanism, where the probability of selecting a unit from a given domain depends on the outcome for that unit. Thus, again, we generate $K = 1,000$ population vectors $\mathbf{y}^{(k)}$, $k = 1, \ldots, K$ from the true nested error model (3.1)-(3.2). For each MC replicate k, we draw a sample $s^{(k)}$. The sample is drawn independently for each domain using Poisson sampling as in the previous experiment. However, in this case the inclusion probability, π_{ij}, for individual j in the sample from domain i depends on a random

Figure 3.1: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under non-informative selection with $\bar{n}_i = 25$.

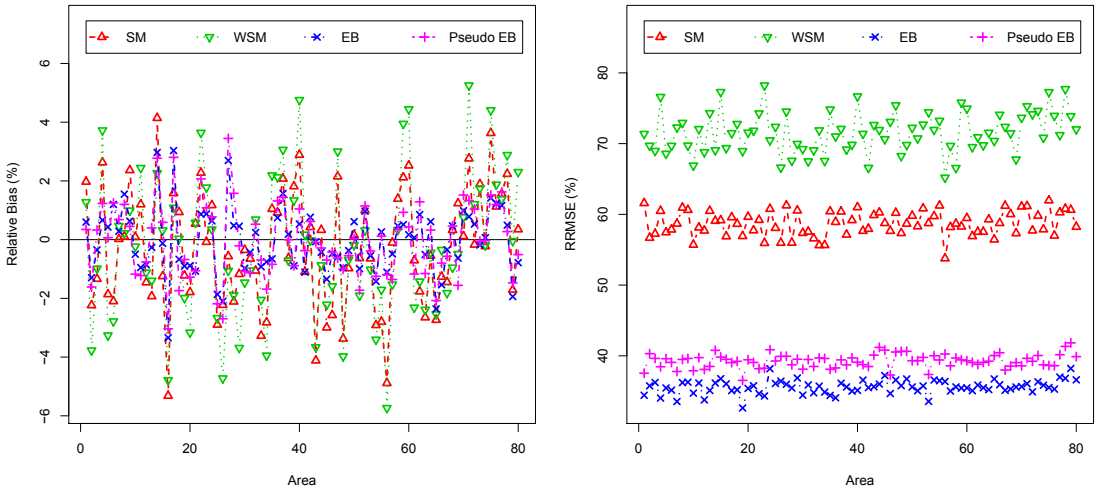


Figure 3.2: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under non-informative selection with $\bar{n}_i = 50$.

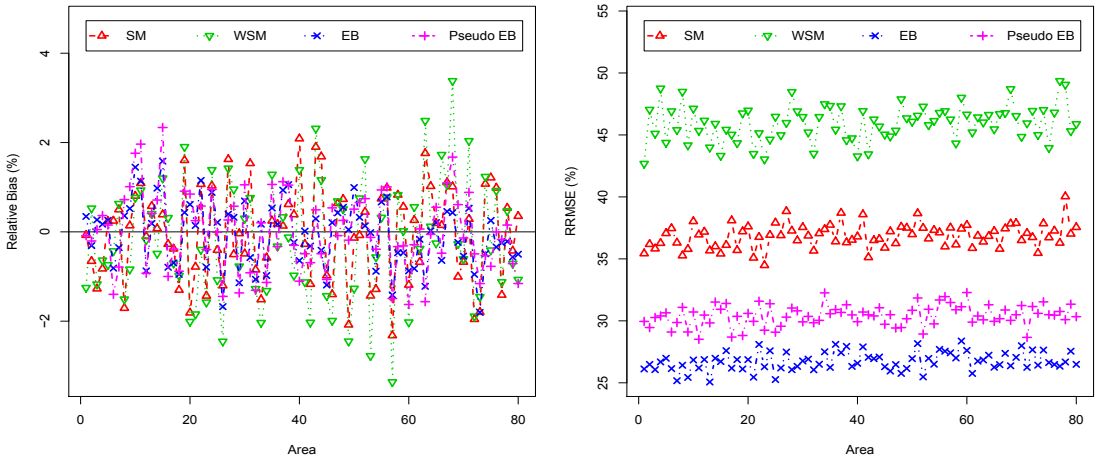
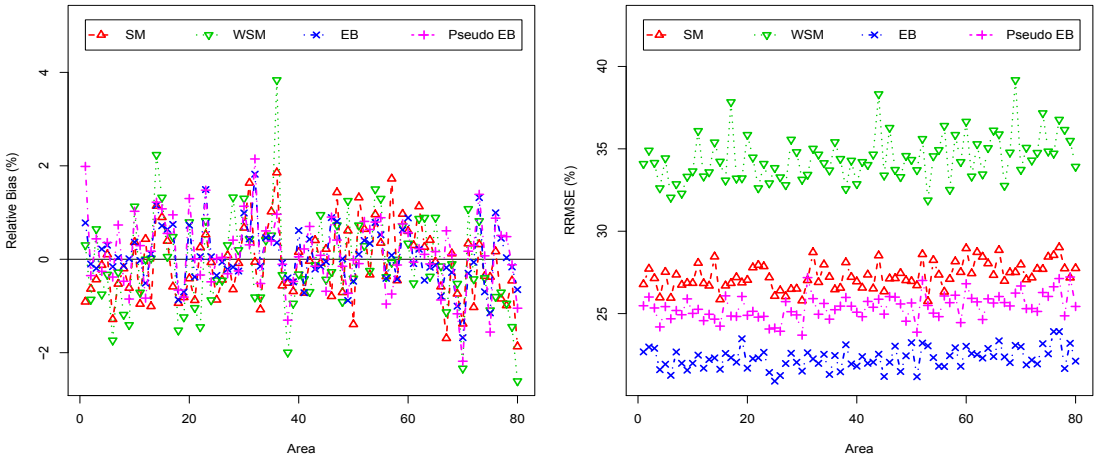


Figure 3.3: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under non-informative selection with $\bar{n}_i = 75$.



variable Z_{ij} that is correlated with the unexplained part of Y_{ij}, i.e, the model error e_{ij}. More concretely, each population unit j comes to the sample s_i from domain i according to a Bernoulli random value $Q_{ij} \sim \text{Bern}(\pi_{ij})$, with $\pi_{ij} = b^{-1}\exp(-aZ_{ij})$, for $a > 0$, $b > 0$, where $Z_{ij} \sim \text{Gamma}(\tau_{ij}, \theta_{ij})$, with model parameters τ_{ij} and θ_{ij} depending on the model error e_{ij}. Here, the degree of informativeness can be measured by the size of the correlation coefficient between Z_{ij} and e_{ij}. An approximately 40% correlation coefficient is achieved by taking $\tau_{ij} = 5 \times (2 + 0.25e_{ij})$ and $\theta_{ij} = 0.25 \times (2 + 0.25e_{ij})$. To make this simulation experiment comparable with the previous one, we take the same expected domain sample sizes $\bar{n}_i = 25, 50, 75$, which can be approximately obtained by fixing $a = 0.15$ and then taking $b = 5.5$ for $\bar{n}_i = 25$, $b = 2.5$ for $\bar{n}_i = 50$ and $b = 1.5$ for $\bar{n}_i = 75$. From each sample $s^{(k)}$, the four estimators (SM, WSM, EB and pseudo EB) are computed.

Figures 3.4, 3.5 and 3.6 depict percent RB (left) and RRMSE (right) under the model and the design of the poverty gap, F_{1i}, for $\bar{n}_i = 25, 50$ and 75 respectively. These figures show how, when the inclusion probabilities are related with the outcome values, the two unweighted estimators (SM and EB) exhibit a substantial positive RB (about 15%). We can see that pseudo EB estimators correct for the strong bias and have smaller RRMSE than EB estimators for all the domains. For the poverty incidence, F_{0i}, plots are not shown because conclusions are similar.

Again, in Table 3.2, we can see average results across domains. This table confirms that the weighted estimators (WSM and pseudo EB) correct for the bias for the three considered expected domain sample sizes, whereas the unweighted estimators (SM and EB) have an average absolute bias over 13% for the poverty incidence, F_{0i}, and over 15%

Figure 3.4: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under informative selection, $\bar{n}_i = 25$.

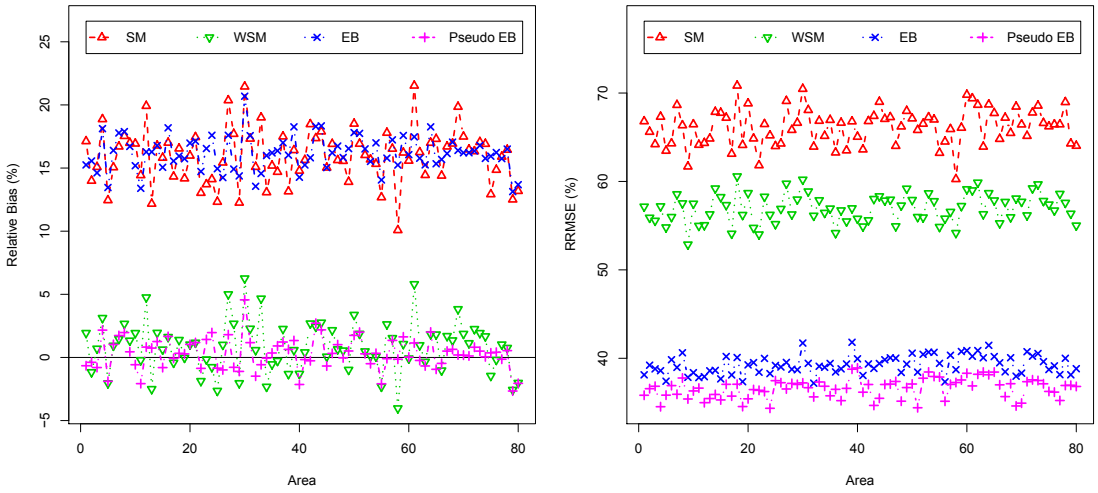


Figure 3.5: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under informative selection, $\bar{n}_i = 50$.

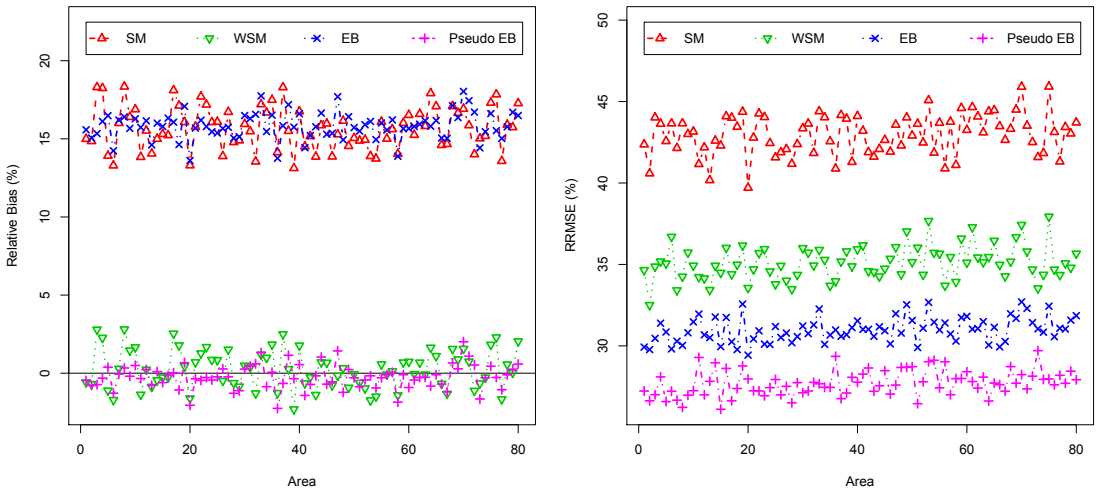
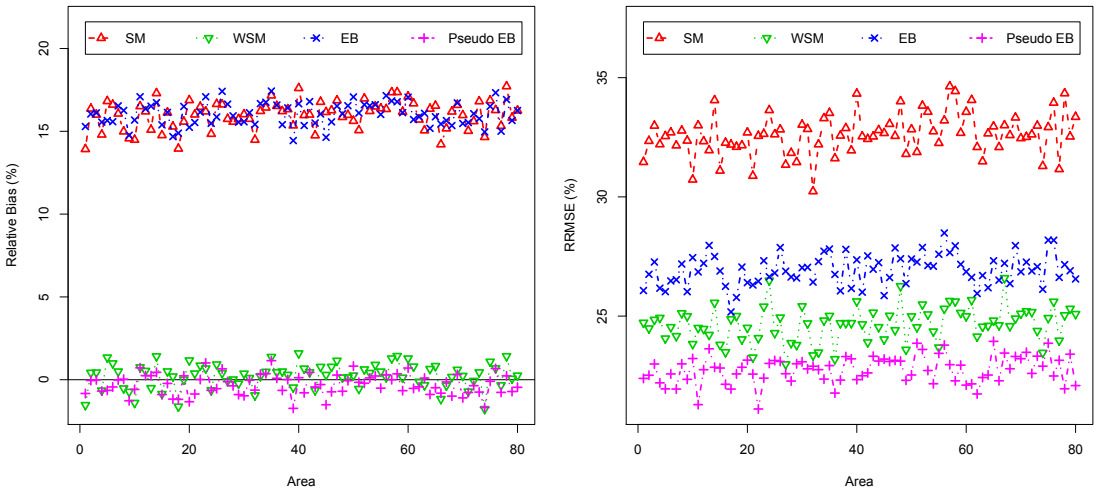


Figure 3.6: Percent RB (left) and RRMSE (right) of SM, WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each area, under informative selection, $\bar{n}_i = 75$.



for the poverty gap, F_{1i}. In terms of average RRMSE, pseudo EB is more efficient than all the other estimators, but the WSM becomes close to the pseudo EB estimator for the largest $\bar{n}_i$. The improvement of the pseudo EB over the unweighted EB estimator in terms of average RRMSE is not striking, but it is in terms of relative bias. Results are in agreement with the results of Pfeffermann and Sverchkov (2007) for estimation of area means under informative sampling.

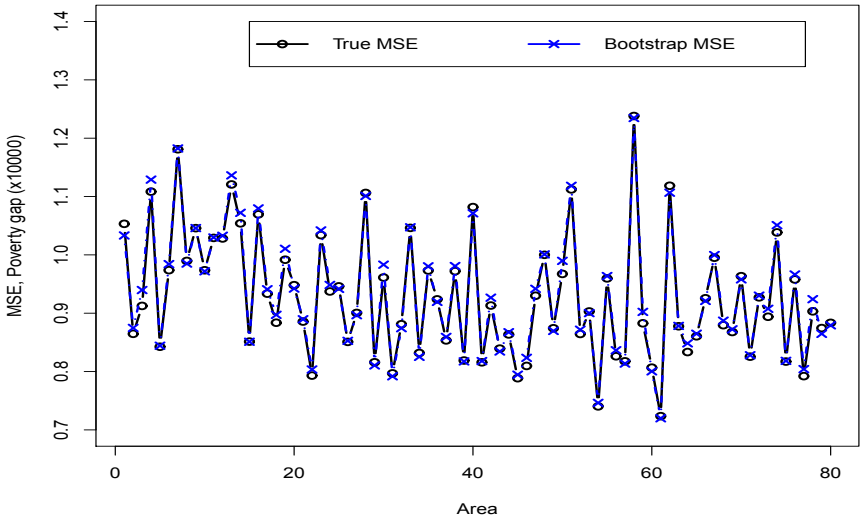
	$\bar{n}_i = 25$				$\bar{n}_i = 50$				$\bar{n}_i = 75$			
	$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$		$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$		$\overline{\text{ARB}}$		$\overline{\text{RRMSE}}$	
Method	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}	F_{0i}	F_{1i}
SM	13.35	15.93	51.14	66.13	13.08	15.66	33.47	42.96	13.12	15.99	25.38	32.61
WSM	1.39	1.72	46.13	56.98	0.83	1.04	28.69	35.11	0.53	0.65	20.15	24.66
EB	13.25	16.15	31.27	39.27	13.09	15.83	24.80	30.98	13.16	16.04	21.53	26.94
PEB	0.79	0.99	29.06	36.59	0.47	0.63	21.94	27.71	0.44	0.55	17.95	22.75

Table 3.2: Averages across domains of percent absolute RB and RRMSE for SM, WSM, EB and pseudo EB estimators of poverty incidence, F_{0i}, and poverty gap, F_{1i}, under informative selection with $\bar{n}_i = 25, 50, 75$.

We also studied the performance of the parametric bootstrap procedure described in Section 3.5 for estimation of the MSE of the pseudo EB estimator. We considered the same simulation setup as above, considering an informative sample, but since the proposed bootstrap procedure gives a model-based MSE, in this case we carry simulations only under the model (given the selected sample). The true MSEs were previously approximated with $K = 10,000$ MC replicates. Then, we perform other

$K = 500$ MC simulation replicates, and in each we calculate the bootstrap MSE estimators (3.18) with $B = 500$ bootstrap replicates. Expected values of the bootstrap MSE estimators across the $K = 500$ MC replicates are shown in Figure 3.7 together with the empirical MSEs for the poverty gap, F_{1i}, with $\bar{n}_i = 50$. This figure shows that the expected values of the bootstrap MSE estimator are almost equal to the true MSE values. Similar results were observed for the poverty incidence, F_{0i}, (not reported).

Figure 3.7: True MSEs of pseudo EB estimators of poverty gap, F_{1i}, and expected values of bootstrap MSE estimators with $B = 500$ bootstrap replicates, for each domain.



3.7 Application to poverty mapping in Mexico

In this section, we apply our proposed method to the estimation of poverty incidences and gaps in the municipalities from the State of Mexico. We use data from the Socioeconomic Conditions Module (SCM) of the 2010 National Survey on Household Income and Expenditure (ENIGH in Spanish). The SCM collects microdata on income, health, nutrition, education, social security, quality of the household, basic household services and social cohesion in Mexico. The SCM sample is drawn independently for each Mexican State and strata and uses two-stage sampling. The primary sample units (PSUs) are groups of adjacent dwellings (between 80 and 300). Sample selection differs for urban, complement urban, and rural places.

The SCM data provides reliable estimates for the overall country disaggregated by urban and rural areas, and for the Mexican States, but it does not allow reliable estimation for municipalities because of the small SCM sample sizes in some of these

domains. In this application, the target areas are the $m = 57$ sampled municipalities (out of 125) in the State of Mexico. We do not intend to estimate in unsampled municipalities because there is no way of checking the model for them. Besides, direct estimates cannot be computed for non-sample areas, then we cannot compare their performance with that of pseudo EB and EB. We excluded municipality coded with 8 because, according to the available information, its sociological structure and economic conditions are very different from the other municipalities. A direct estimator can be given for this municipality. The overall sample size is 10,560 after removing records with missing values, and the population size is 14,670,655.

As auxiliary variables in the nested error model, we consider age, squared age and age to the third power, since a scatterplot of income against age displays a polynomial pattern. We also include the indicators of gender, indigenous community, receiving government benefits, of rural or urban location, six levels of activity sector (first, second and third sector, unemployed, inactive and below working age), four levels of household quality depending on household conditions and services, three levels of household structure (single adult, primary family group and extended families), years of study recoded in four groups (no-studies, primary, secondary and university studies) and, finally, the variable household services with four groups depending on how many services from a given list, the household members have access to.

The considered welfare measure E_{ij} is the monthly total current per capita income (tcpci). The poverty line is established by the National Council for the Evaluation of the Social Development Policy (CONEVAL in Spanish) and it varies depending on the intensity of the poverty that we wish to measure (moderate or extreme) and for rural or urban places. We estimate here moderate poverty in rural and urban places. For that year, the poverty line is \$2,113.86 for urban places and \$1,328.51 for rural places. Finally, since income has a fairly skewed distribution, we transform it like in Molina et al. (2014), by $Y_{ij} = T(E_{ij}) = \log(E_{ij} + c)$, where c is selected such that the distribution of the residuals obtained from the model fit, $\hat{e}_{ij} = Y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} - \hat{v}_i$, is approximately symmetric.

We wish to compare EB and pseudo EB estimates and their estimated coefficients of variation (in other words, estimated RRMSEs). Instead of the original EB and pseudo EB methods, since the sampling fractions are very small for all the municipalities, we applied Census EB and Census PEB respectively.

Before comparing these estimates, let us first analyze whether model assumptions hold. Note that a sample under informative sampling need not follow the same population model. Here we assume that, even if informative sampling can change the model parameters, the shape of the model for sample elements is the same as for

the population. In this case, using the sampling weights to fit the model corrects for informative sampling. Moreover, if unweighted and weighted fitted values are similar, we can suspect that informative sampling is not having an effect on the fitted model. Thus, first we look at residuals obtained from the unweighted fit of the model $\hat{e}_{ij} = Y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} - \hat{v}_i$ (called hereafter EB residuals) against predicted values $\hat{Y}_{ij} = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \hat{v}_i$ in Figure 3.8 (left). Figure 3.8 (right) shows the analogous plot for the residuals obtained from the weighted fit, $\hat{e}_{ijw} = Y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}_w - \hat{v}_{iw}$ (called pseudo EB residuals), against the corresponding predicted values $\hat{Y}_{ijw} = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}_w + \hat{v}_{iw}$. These two plots look pretty similar, so it seems that the effect of informative sampling is in this case small. The points that appear aligned are for residuals corresponding to the same income value (incomes are integer values and some of them are repeated several times in the sample). Apart from this fact, none of the plots show model departures. Normal Q-Q plots of EB and pseudo EB residuals included in Figure 3.9 are also pretty similar, showing that the distributions of EB and pseudo EB residuals have both slightly heavier tails than the normal distribution. Figure 3.10 shows normal Q-Q plots of unweighted estimates of area effects $\hat{v}_i$ (left) and weighted ones $\hat{v}_{iw}$ (right) for each sampled municipality. In both cases, the distribution of estimated area effects is similar to a normal distribution.

Now, to check whether the sampling design is actually informative and we should then use estimators that account for the sampling weights, we first compare unweighted direct estimates (sample means) with weighted direct estimates (weighted sampled means). Figure 3.11 plots unweighted direct estimates of poverty incidences (left) and poverty gaps (right) against weighted direct ones for each municipality. Those municipalities corresponding to points lying exactly on the line have constant sampling weights for all their sampled individuals, whereas those few whose points appear to be further from the line (highlighted in red) have unequal sampling weights, such that weighted and unweighted direct estimates clearly differ.

Let us now compare model-based estimates (EB and pseudo EB) with weighted direct ones. Figure 3.12 displays EB estimates (left) and pseudo EB estimates (right) of poverty incidences for each municipality against weighted direct estimates, with municipality codes as point labels. On the right plot, points seem to be more spread along the line than on the left plot. Since weighted direct estimates (on the x-axis) are design-unbiased, this might suggest a slightly better performance of pseudo EB estimates compared to EB estimates in terms of design-bias.

Since informativeness can only affect municipalities for which weighted and unweighted estimates differ, in order to gain more insight into the effects of sample informativeness, we repeat the overall study (model selection, model fitting and estimation) only for those municipalities where weighted and unweighted direct estimates differ

Figure 3.8: EB (left) and pseudo EB (right) residuals against predicted values.

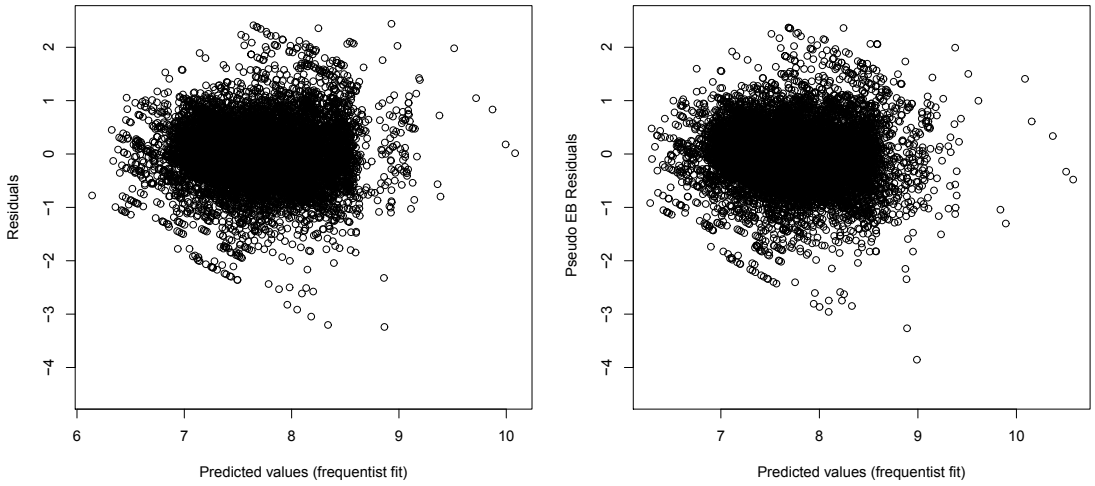


Figure 3.9: Normal Q-Q plots of EB (left) residuals and pseudo EB (right) residuals.

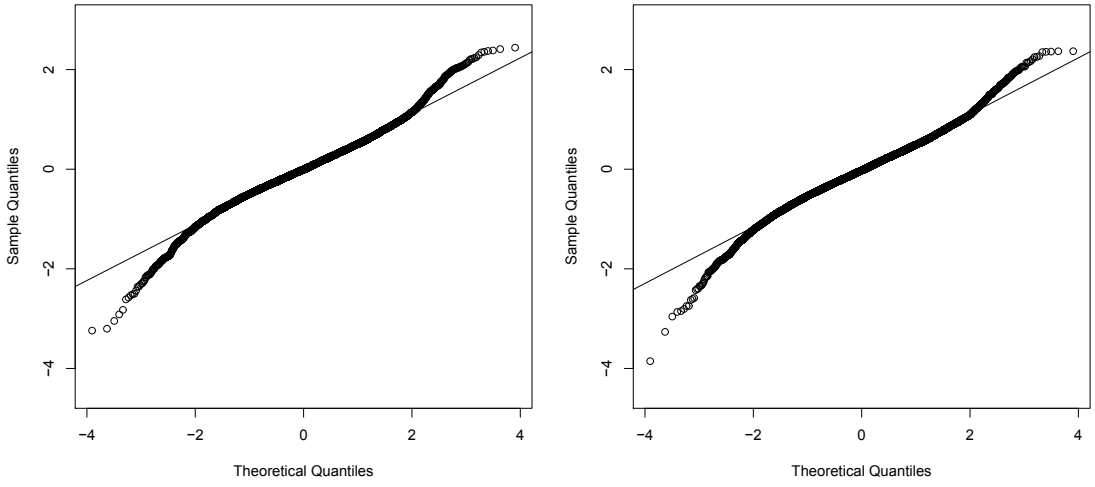
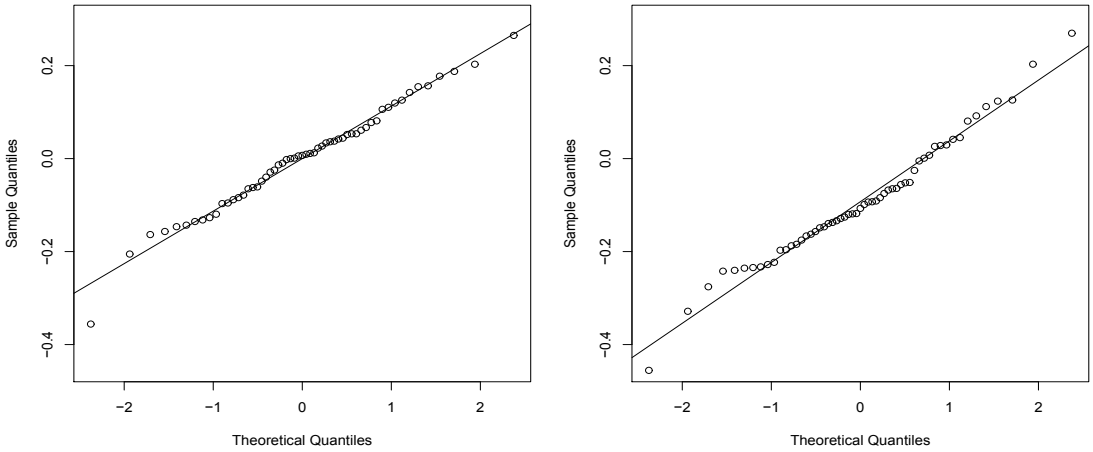


Figure 3.10: Normal Q-Q plot of EB (left) and pseudo EB (right) predicted municipality effects.



(highlighted in Figure 3.11). The resulting model is slightly different for these selected municipalities because some of the auxiliary variables considered before are no longer significant. To check this new model for the reduced sample, let us look at Figure 3.13, showing EB residuals $\hat{e}_{ij}$ (left) and pseudo EB residuals $\hat{e}_{ijw}$ (right) against predicted values $\hat{Y}_{ij}$ and $\hat{Y}_{ijw}$ respectively. Again, these plots show no clear model departures. Similarly as before, normal Q-Q plots given in Figure 3.14 show distributions of EB and pseudo EB residuals with slightly heavier tails than the normal distribution. Normal Q-Q plots of unweighted predictors of area effects $\hat{v}_i$ (left) and weighted predictors $\hat{v}_{iw}$ (right) for each selected municipality show no departure from normality as can be seen in Figure 3.15.

We now study sample ignorability (or prediction bias) for the selected municipalities using the test proposed by Pfeffermann and Sverchkov (2007). This test checks whether the regression coefficient for the study variable (in our case income) is significant when we regress sampling weights against the auxiliary variables and the study variable for each municipality. Table 3.3 reports the observed F-values for the coefficients of our study variable for each selected municipality. In that table, the maximum F-value, $F_{\max} = 147.81$, is much larger than the corresponding 0.05-critical point $F_{\max,0.05} = 11.97$ of $F_{\max}$, which was obtained by MC simulation. In fact, the table shows F-values for three municipalities larger than $F_{\max,0.05} = 11.97$. Then, we conclude that there is at least one of the municipalities where the regression coefficient of the study variable differs significantly from zero; therefore, the sampling design should not be ignored here.

Let us now compare model-based estimators (EB and pseudo EB) against weighted direct estimators (WSMs) for the selected municipalities. Figure 3.16 plots EB estimates

Municipality	9	13	20	37	39	51	54	95	103	114	118	121
n_i	213	291	120	43	124	234	469	145	131	279	182	227
F-value	4.43	147.81	6.01	0.183	2.59	8.15	0.28	11.17	0.14	2.56	20.56	15.52

Table 3.3: Test for sample ignorability. Code, sample size and observed F-value for each selected municipality.

(left) and pseudo EB estimates (right) of poverty incidence for each municipality against WSMs. Results look similar to those obtained previously with points slightly more spread along the line for pseudo EB.

Figures 3.17 and 3.18 report the resulting estimates and the estimated coefficients of variation (CVs) for selected municipalities, obtained as estimated root MSE by the corresponding estimate (in %). Since the considered direct estimators (WSMs) are ratio estimators, their MSE was calculated using the Taylor linearization method. For EB estimators, the MSE was obtained using the parametric bootstrap approach of Molina and Rao (2010). For pseudo EB estimators, the MSE was approximated by the bootstrap procedure of Section 3.5. Figures 3.17 (left) and 3.18 (left) show that direct estimates are more unstable than pseudo EB and EB. Moreover, in terms of efficiency, direct estimates get very large estimated CVs for municipalities with smaller sample sizes. As expected, estimated CVs of pseudo EB estimators are slightly larger than those of EB estimators except for the municipalities with the smallest sample sizes. This is the price to pay for reducing the design bias. Nevertheless, pseudo EB estimators lead to large reduction in CV relative to direct estimators for all but one area.

Figure 3.19 displays cartograms of direct (top left), EB (top right) and pseudo EB (bottom left) estimates of poverty incidence F_{0i} for each of the selected municipalities. Figure 3.20 shows the analogous estimates for the poverty gap. We considered different poverty intervals and colors for each method because the ranges of direct, EB and pseudo EB estimates differ quite a lot. These figures indicate that the largest poverty incidences and gaps are for the selected municipalities around the center of the State of Mexico. EB estimates provide a larger number of municipalities with poverty incidence in the last interval of extreme poverty than pseudo EB ones. We can see colors also tending to be darker for EB estimates than for pseudo EB ones in the case of poverty gap. This might be a sign of the overestimation of EB method that we also observed in our simulation results, and which seems to be corrected by our pseudo EB method.

Figure 3.11: Unweighted direct estimates of poverty incidence (left) and poverty gap (right) against weighted direct estimates for each sampled municipality.

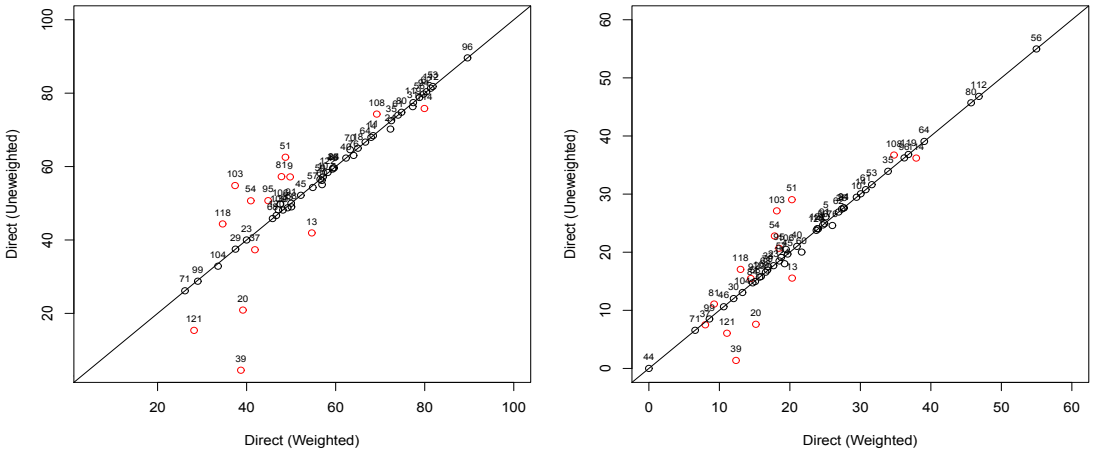


Figure 3.12: EB estimates of poverty incidence (left) and pseudo EB (right) against weighted direct estimates for each sampled municipality.

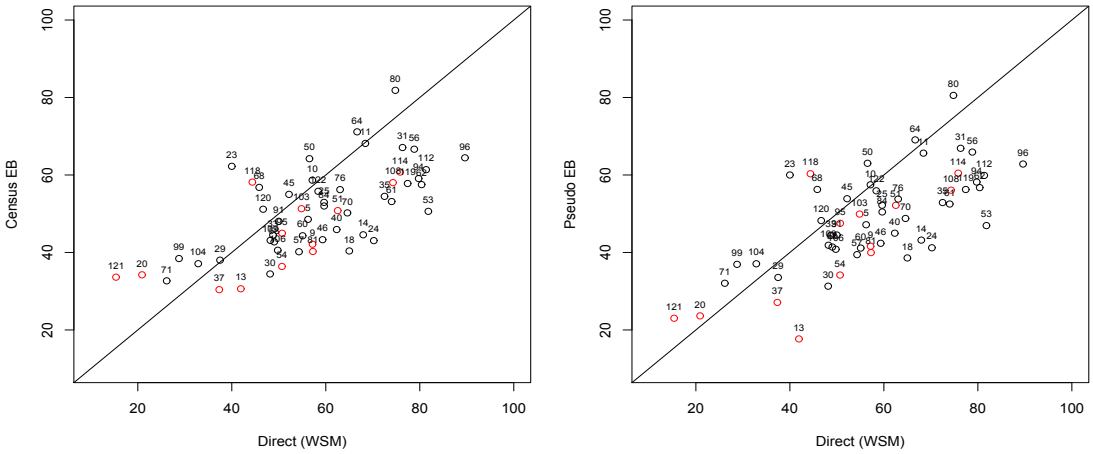


Figure 3.13: EB (left) and pseudo EB (right) residuals against predicted values for selected municipalities.

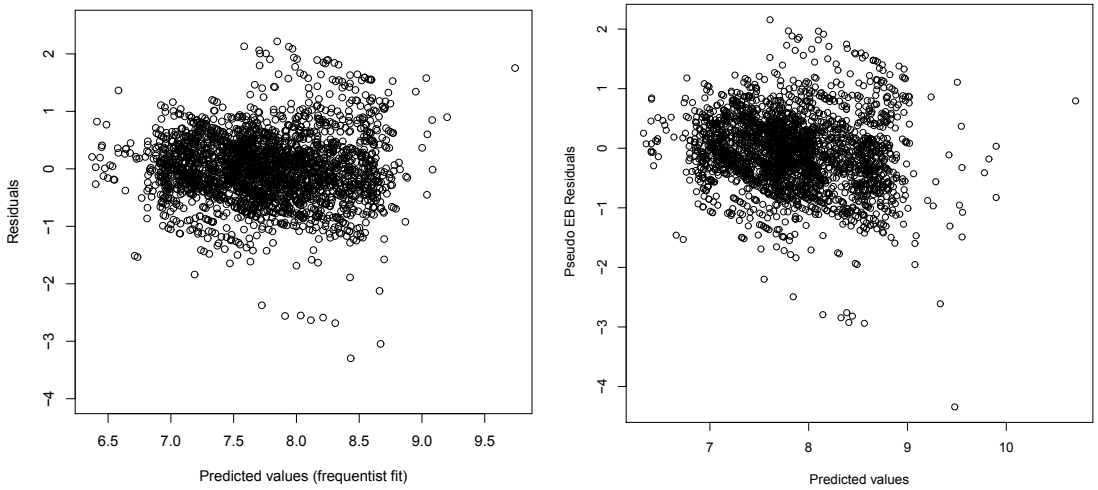


Figure 3.14: Normal Q-Q plot of EB residuals (left) and pseudo EB residuals (right) for selected municipalities.

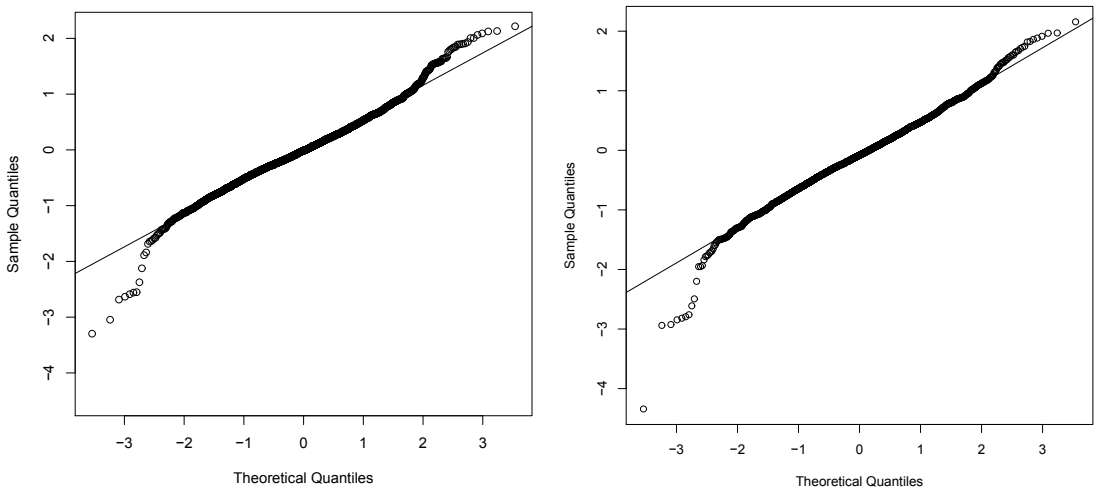


Figure 3.15: Normal Q-Q plot of estimated effects by EB (left) and pseudo EB (right) for each sampled municipality i.

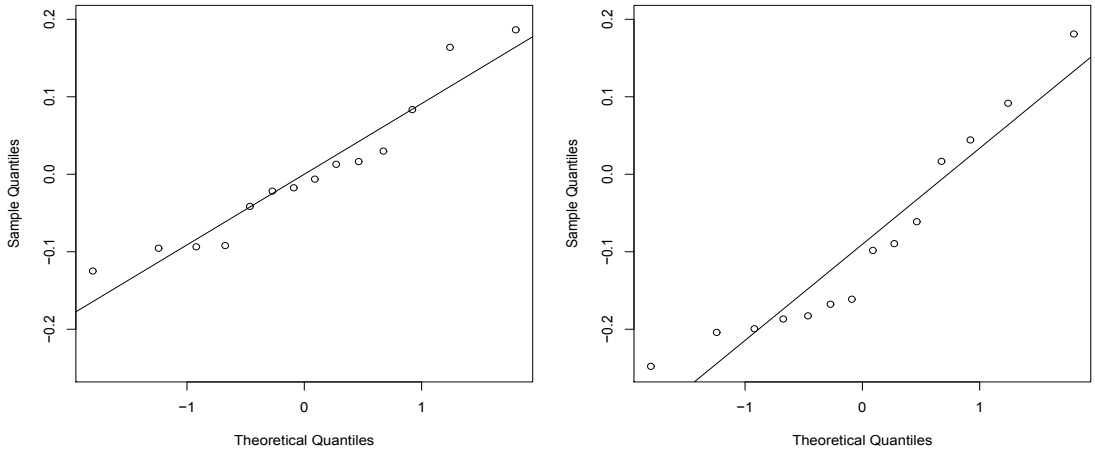


Figure 3.16: EB estimates of poverty incidence (left) and pseudo EB (right) against weighted direct estimates for selected municipalites.

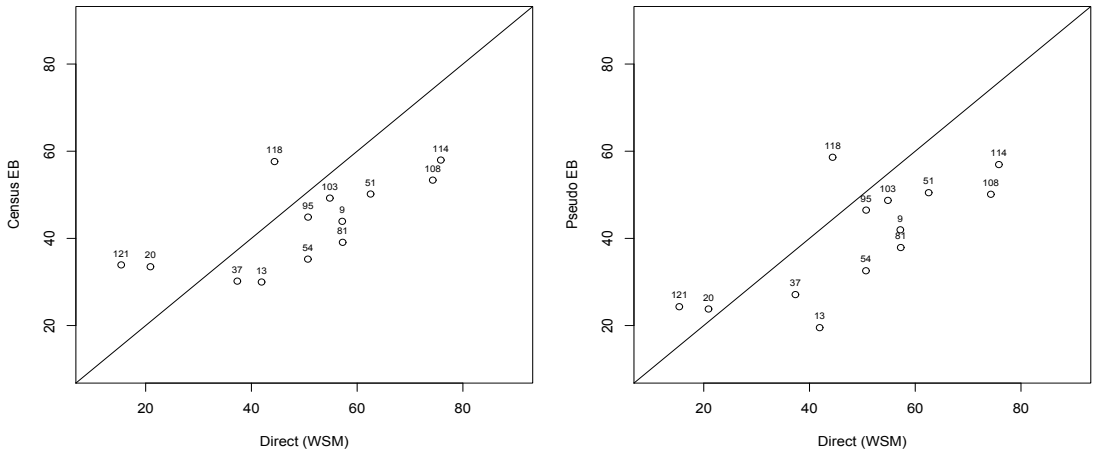


Figure 3.17: Estimates (left) and coefficients of variation (right) of WSM, EB and pseudo EB estimators of poverty incidence, F_{0i}, for each selected municipality.

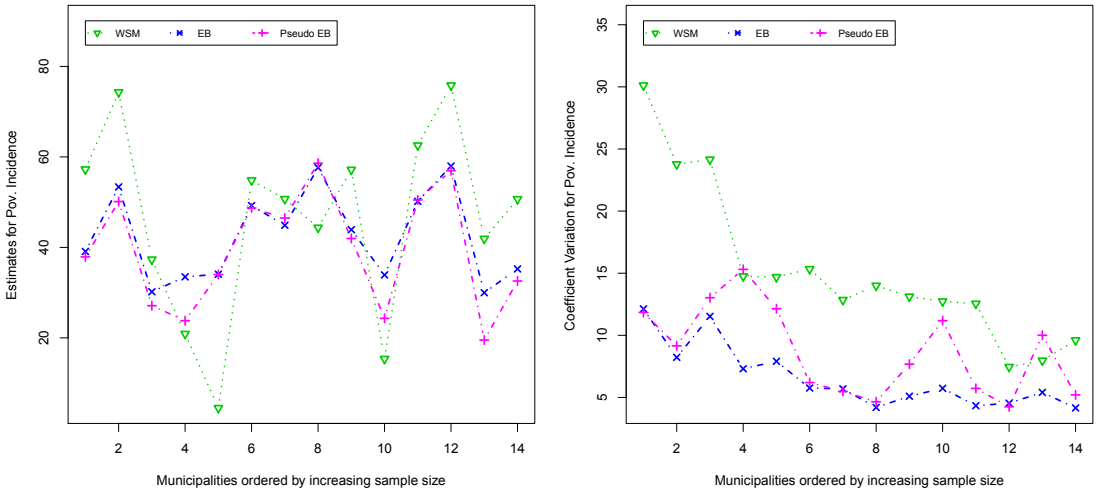


Figure 3.18: Estimates (left) and coefficient of variation (right) of WSM, EB and pseudo EB estimators of poverty gap, F_{1i}, for each selected municipality.

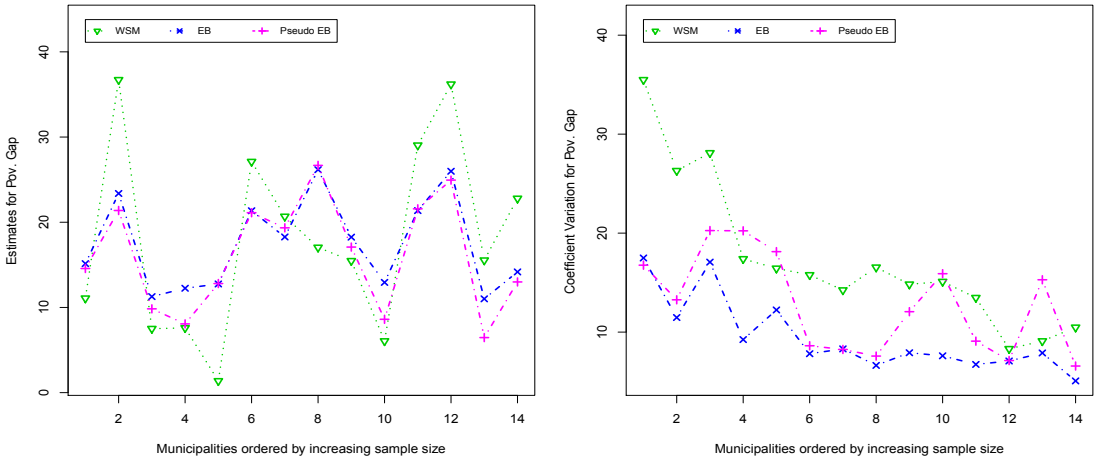
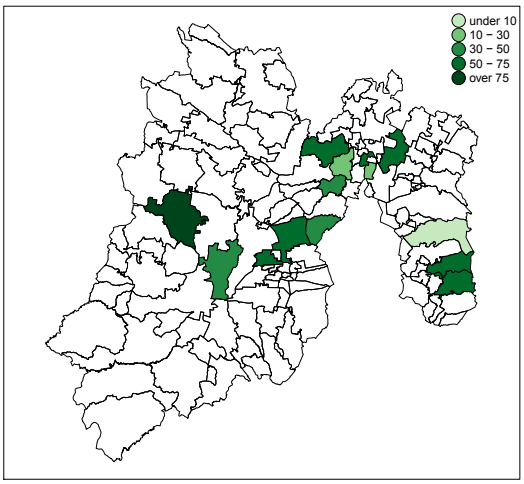
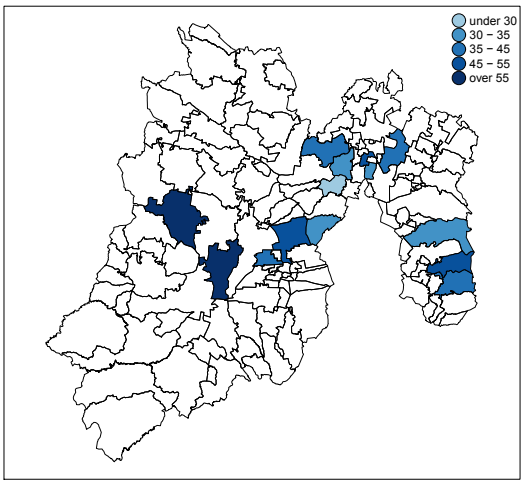


Figure 3.19: Cartograms of estimated percent poverty incidences, F_{0i}, in the selected municipalities from the State of Mexico, obtained with direct (top left), EB (top right) and pseudo EB (bottom left) methods.





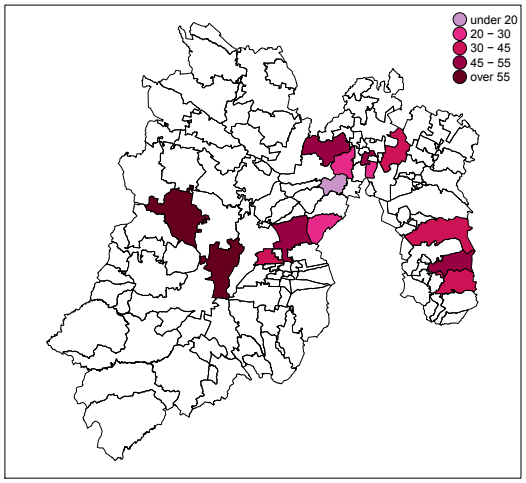
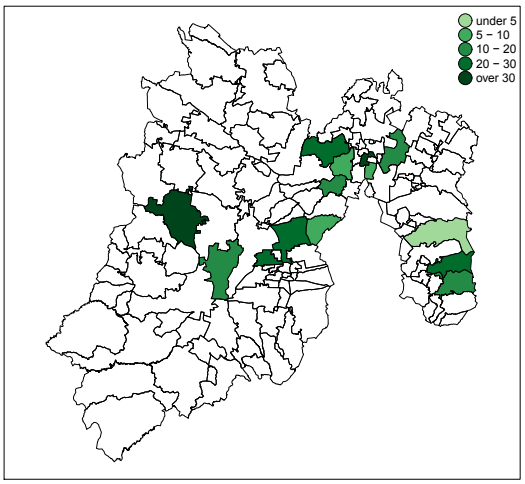
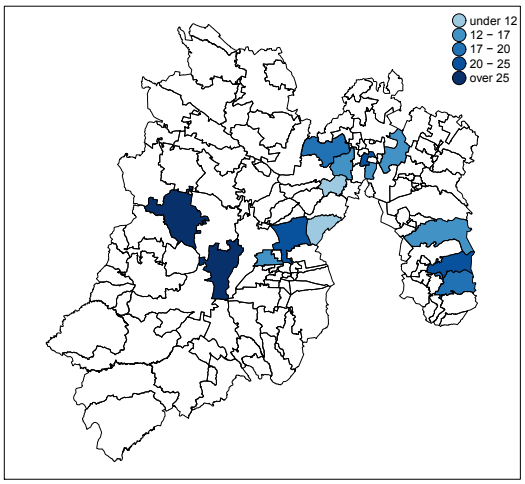
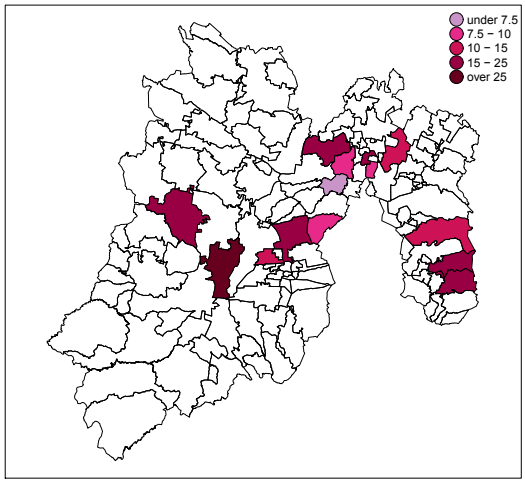


Figure 3.20: Cartograms of estimated percent poverty incidences, F_{1i}, in the selected municipalities from the State of Mexico, obtained with direct (top left), EB (top right) and pseudo EB (bottom left) methods.







Chapter 4

Small area estimation methods under cut-off sampling

The OECD defines cut-off sampling as a sampling procedure in which a predetermined threshold is established with all units in the population and all units at or above (below) the threshold are excluded from the possible selection in a sample. This procedure leads to biased estimates since the inclusion probabilites for excluded units are zero, see e.g. Särndal et al. (1992), Haziza et al. (2010) among others. Haziza et al. (2010) propose to use auxiliary information in order to reduce the bias when estimating population totals at the estimation stage; more concretely, they propose to use balanced sampling and/or calibration. In this chapter, we study how cut-off sampling affects the estimation of domain (or area) parameters and study some of the calibration methods proposed by Haziza et al. (2010) to reduce this problem. For domains with small sample size (small domains or areas), calibration estimators might suffer from large sampling variances. For this case, we will study the empirical best linear unbiased predictor (EBLUP). We apply the methods proposed in this chapter to the estimation of the total sales of certain tobacco product in the provinces from Spain.

The material is organized as follows. In Section 4.1, we start reviewing the basic design-based estimators, as well as calibration estimators, when estimating at the national level, switching in Section 4.2 to estimation at the domain (or area) level, in absence of cut-off sampling. Section 4.3 then focuses on the estimation methods proposed to handle cut-off sampling. This section describes the basic small area estimation methods, namely the EBLUP for estimation of linear parameters, studying its properties under cut-off sampling, and the EB method for estimation of general parameters. Section 4.4 describes a bootstrap procedure for estimation of the model mean squared error. Section 4.5 compares the performance of the considered calibration

estimators and of the small area estimators under cut-off sampling through a simulation study. Section 4.6 describes an application to the estimation of the total sales of a certain tobacco product in Spanish provinces.

4.1 Estimation of population totals or means

Design-based estimators have good properties under the sample replication mechanism, which assumes that the values of the target variable in the population units are fixed. In this section, we describe basic design-based estimators as well as calibration estimators that use auxiliary information, under complex designs but with strictly positive inclusion probabilities for all population units.

4.1.1 Basic design-based estimators

We start describing two basic design-based estimators, namely the well-known Horvitz-Thompson and Hájek estimators. Let U be a finite population of size N. We denote by y_j the study variable for j-th population unit, assumed to be fixed. The target parameter is the population total, $Y = \sum_{j=1}^{N} y_j$. To estimate Y, we draw a sample s of size n from the population U. Let $\pi_j = \Pr(j \in s) > 0$ be the inclusion probability for unit $j \in U$ in the sample and $w_j = \pi_j^{-1}$ the corresponding sampling weight. The Horvitz-Thompson (HT) estimator, also called "expansion" estimator, was proposed by Horvitz and Thompson (1952). For a population total, $Y = \sum_{j=1}^{N} y_j$, and for the corresponding population mean, $\bar{Y} = N^{-1} \sum_{j=1}^{N} y_j$, HT estimators are given respectively by

$$\hat{Y} = \sum_{j \in s} w_j y_j, \qquad \hat{\bar{Y}} = N^{-1} \sum_{j \in s} w_j y_j. \tag{4.1}$$

HT estimator is design-unbiased and design-consistent as the sample size n increases even under complex sampling designs, as long as it is calculated using the correct inclusion probabilities.

Hájek (HA) estimator, proposed by Hájek (1971), is defined in terms of the weighted sample mean (WSM), using as weights those coming from the sampling design. For the total, Y, and the mean, $\bar{Y}$, the Hájek estimators are respectively given by

$$\hat{Y}^{HA} = N\hat{\bar{Y}}/\hat{N}, \qquad \hat{\bar{Y}}^{HA} = \hat{Y}/\hat{N}, \tag{4.2}$$

where $\hat{N} = \sum_{j \in s} w_j$ is the HT estimator of N. Hájek estimator is nearly design-unbiased and also design consistent as n increases for general sampling designs.

When the sample is drawn by simple random sampling without replacement (srswor), we have $w_j = N/n,\ \forall j \in U$ and $\sum_{j\in s} w_j = N$. In this case, HT and HA estimators of the mean, $\bar{Y}$, reduce to the usual sample mean $\hat{\bar{Y}} = \hat{\bar{Y}}^{HA} = \bar{y}_s$ and, for the total, they reduce to $\hat{Y} = \hat{Y}^{HA} = N\bar{y}_s$.

4.1.2 Calibration estimators

Calibration was first proposed by Deville and Särndal (1992) to estimate characteristics of the study variable, y, employing auxiliary information. Let $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})'$, be a vector of p auxiliary variables available for all units in the sample with known vector of population totals $\mathbf{X} = \sum_{j=1}^{N} \mathbf{x}_j$.

The idea of calibration estimators is that, if we estimate correctly the total or mean of $\mathbf{x}_j$ and $\mathbf{x}_j$ is linearly related with y_j, we will estimate correctly also the total or mean of y_j. The aim of calibration is to obtain new weights $h_j,\ j \in s$, that are as close as possible to the original sampling weights, $w_j,\ j \in s$, and which have the desiderable property of yielding exact estimators of the totals for the auxiliary variables, in the sense of satisfying the calibration equations

$$\sum_{j\in s} h_j \mathbf{x}_j = \mathbf{X}. \tag{4.3}$$

Let $G_j(h, w)$ be a distance measure between the new weight h_j and the original weight, w_j; for example, the chi-squared distance given by $G_j(h, w) = (h_j - w_j)^2/w_j$. Calibration solves the following problem to obtain the new weights $h_j,\ j \in s$:

$$\begin{aligned} \min \quad & \sum_{j\in s} G_j(h, w) \\ \text{s.t.} \quad & \sum_{j\in s} h_j \mathbf{x}_j = \mathbf{X}. \end{aligned}$$

In the case of the chi-squared distance $G_j(h, w) = (h_j - w_j)^2/w_j$, using the method of Lagrange multipliers, the Lagrangian function is given by

$$L = \sum_{j\in s} (h_j - w_j)^2/w_j + 2\boldsymbol{\lambda}' \left(\sum_{j\in s} h_j \mathbf{x}_j - \mathbf{X} \right),$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. Taking derivatives of L with respect to the new weights h_j and $\boldsymbol{\lambda}$ and equating to zero we obtain the solution to the above

problem given by

$$h_j = w_j(1+\mathbf{x}_j'\boldsymbol{\lambda}),\ j\in s \tag{4.4}$$

$$\boldsymbol{\lambda} = \left(\sum_{j\in s} w_j\mathbf{x}_j\mathbf{x}_j'\right)^{-1}(\mathbf{X}-\hat{\mathbf{X}}), \tag{4.5}$$

where $\hat{\mathbf{X}} = \sum_{j\in s} w_j\mathbf{x}_j$ is the usual expansion (or HT) estimator of $\mathbf{X}$. The calibration estimator obtained from the chi-squared distance, called linear calibration estimator (LCAL), is then obtained using the new weights as

$$\hat{Y}^{LCAL} = \sum_{j\in s} h_j y_j. \tag{4.6}$$

Deville and Särndal (1992) noticed that the calibration estimator (4.6) of the total Y obtained by minimizing the chi-squared distance equals the well known generalized regression (GREG) estimator, defined as

$$\hat{Y}^{GREG} = \hat{Y} + (\mathbf{X}-\hat{\mathbf{X}})'\hat{\mathbf{B}} \tag{4.7}$$

where $\hat{Y} = \sum_{j\in s} w_j y_j$ and $\hat{\mathbf{B}} = (\sum_{j\in s} w_j\mathbf{x}_j\mathbf{x}_j')^{-1}\sum_{j\in s} w_j\mathbf{x}_j y_j$, provided that $\sum_{j\in s} w_j\mathbf{x}_j\mathbf{x}_j'$ is non-singular, see Appendix C.1 for the proof. The GREG estimator is motivated by a linear regression model for y_j in terms of $\mathbf{x}_j$ of the form

$$y_j = \mathbf{x}_j'\boldsymbol{\beta} + \epsilon_j,\ j\in s, \tag{4.8}$$

where ϵ_j is the model error with $E_m(\epsilon_j)=0$, where E_m denotes expectation under the model (4.8), $E_m(\epsilon_j,\epsilon_\ell)=0$ for $j\neq\ell$ and variance $V_m(\epsilon_j)=\sigma_\epsilon^2$. We do not consider heteroscedasticity for simplicity. Note that $\hat{\mathbf{B}}$ is the weighted least squared estimator (WLS) estimator of $\boldsymbol{\beta}$ under this model obtained using the sampling weights, which is the sample version of the population regression parameter $\mathbf{B} = (\sum_{j=1}^N \mathbf{x}_j\mathbf{x}_j')^{-1}\sum_{j=1}^N \mathbf{x}_j y_j$. However, the GREG (4.7), or LCAL estimator, has good properties with respect to the design even if the model (4.8) fails. Specifically, it is design consistent as the sample size, n, increases; in this sense, the GREG is called a model-assisted estimator.

The chi-squared distance provides sometimes negative weights h_j and, in some cases, these negative weights may have no sense. Deville and Särndal (1992) proposed alternative distance measures, $G_j(h,w)$, that satisfy two conditions. The first one is that for $w_j>0$, $G_j(h,w)$ is non-negative. The second one is that the first derivative of $G_j(h,w)$, viewed as a function of h, is increasing and continuous. These two conditions ensure that, if a solution to the calibration problem exists, it is unique, and the new

weights can always be written as

$$h_j = w_j F(\mathbf{x}_j'\boldsymbol{\lambda}), \tag{4.9}$$

where F is the inverse of the first derivative of $G_j(h, w)$ with respect to h divided by w, seen as a function of h, see Appendix C.2 for derivation of $F(\cdot)$. Note that for the chi-squared distance, $G_j(h, w) = (h_j - w_j)^2/w_j$, the new weights in (4.4), have the form (4.9) with $F(\mathbf{x}_j'\boldsymbol{\lambda}) = 1 + \mathbf{x}_j'\boldsymbol{\lambda}$, which is a linear function of $\mathbf{x}_j'\boldsymbol{\lambda}$. This fact gives the name linear calibration estimator (LCAL) to $\hat{Y}^{LCAL}$ in (4.6).

Once $F(\cdot)$ is determined for a given distance function $G_j(h, w)$ and $\boldsymbol{\lambda}$ is obtained, the calibration estimator of the total Y is then

$$\hat{Y}^{CAL} = \sum_{j \in s} h_j y_j = \sum_{j \in s} w_j F(\mathbf{x}_j'\boldsymbol{\lambda}) y_j. \tag{4.10}$$

Deville and Särndal (1992) prove that for any $F(\cdot)$ satisfying certain regularity conditions, $\hat{Y}^{CAL}$ is asymptotically equivalent to the GREG estimator $\hat{Y}^{GREG}$. Thus, under those conditions, the two estimators share the same asymptotic variance. The efficiencies of the GREG and the calibration estimators as compared to the basic design-based estimators of Section 4.1.1 depend on the goodness of fit of the regression model given in (4.8), see Särndal et al. (1992).

4.2 Domain estimators

Now we focus on the estimation of totals or means for domains of the population. We will first describe the basic direct estimators, which typically use only the data of the study variable from the corresponding domain. After that, we describe calibration estimators for domain totals/means, which use auxiliary information. Thus, in the following, we assume that the population U is divided into m non-overlapping subsets U_i, $i = 1, \ldots, m$, which are the domains or areas, of sizes N_i, $i = 1, \ldots, m$, with $N = \sum_{i=1}^{m} N_i$. Hereafter, we use subscript j for the unit within a domain and i for the domain and we denote by y_{ij} the study variable for j-th unit within i-th domain.

4.2.1 Basic direct estimators

We may now estimate the total $Y_i = \sum_{j=1}^{N_i} y_{ij}$ or the mean $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$ for each domain i. Then, we draw a sample s_i of size n_i from domain U_i, $i = 1, \ldots, m$, and define $\pi_{ij} = \Pr(j \in s_i) > 0$ as the inclusion probability of unit $j \in U_i$ in the sample s_i from domain i and $w_{ij} = \pi_{ij}^{-1}$ the corresponding sampling weight for the same unit.

As already mentioned, the basic direct estimators use only the n_i observations of the variable of interest from area i and HT and HA estimators are good examples. The HT or expansion estimators of the domain total Y_i and the domain mean $\bar{Y}_i$ are given respectively by

$$\hat{Y}_i = \sum_{j \in s_i} w_{ij} y_{ij}, \qquad \hat{\bar{Y}}_i = {N_i}^{-1} \sum_{j \in s_i} w_{ij} y_{ij}, \tag{4.11}$$

whereas HA estimators are respectively given by

$$\hat{Y}_i^{HA} = N_i \hat{\bar{Y}}_i / \hat{N}_i, \qquad \hat{\bar{Y}}_i^{HA} = \hat{Y}_i / \hat{N}_i \tag{4.12}$$

Again, under srswor within the domain U_i, we have $w_{ij} = N_i/n_i,\ \forall j \in U_i$, and then $\sum_{j \in s_i} w_{ij} = N_i$. In this case, HT and HA estimators of $\bar{Y}_i$ reduce both to the sample mean from area i, $\hat{\bar{Y}}_i = \hat{\bar{Y}}_i^{HA} = \bar{y}_{is}$ and for the total they become $\hat{Y}_i = \hat{Y}_i^{HA} = N_i \bar{y}_{is}$.

In the case of estimation in domains, direct estimators are design consistent as the domain sample size n_i increases. However, they are very inefficient for domains with very small sample sizes. In fact, in the case that not all the domains are sampled, they cannot be calculated for nonsampled domains (i.e., with $n_i = 0$).

4.2.2 Calibration estimators

Let $\mathbf{x}_{ij}$ be a vector with the values of p auxiliary variables for the unit j within domain i. We can use different approaches to obtain a calibration estimator for the total Y_i in domain i. Using as an example the chi-squared distance, one approach would be to minimize the sum of distances, $G_{ij}(h, w) = (h_{ij} - w_{ij})^2 / w_{ij}$ for the sample units in that domain, subject to the calibration equations for the same domain i. In this case, the calibration problem is

$$\begin{aligned} \min \quad & \sum_{j \in s_i} (h_{ij} - w_{ij})^2 / w_{ij} \\ \text{s.t.} \quad & \sum_{j \in s_i} h_{ij} \mathbf{x}_{ij} = \mathbf{X}_i, \end{aligned} \tag{4.13}$$

where the true domain total $\mathbf{X}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ is assumed to be known. Solving (4.13) by the method of Lagrange multipliers, where

$$L_i = \sum_{j \in s_i} (h_{ij} - w_{ij})^2 / w_{ij} + 2 \boldsymbol{\lambda}_i' \left(\sum_{j \in s_i} h_{ij} \mathbf{x}_{ij} - \mathbf{X}_i \right),$$

is the lagrangian function and $\boldsymbol{\lambda}_i$ is the vector of Lagrange multipliers for area i, we obtain

$$\begin{aligned} h_{ij} &= w_{ij}(1+\mathbf{x}'_{ij}\boldsymbol{\lambda}_i),\ j \in s_i, \\ \boldsymbol{\lambda}_i &= \left(\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}'_{ij}\right)^{-1}(\mathbf{X}_i - \hat{\mathbf{X}}_i). \end{aligned} \tag{4.14}$$

The calibration estimator of the domain total Y_i is then given by

$$\hat{Y}_i^{LCAL} = \sum_{j\in s_i} h_{ij}y_{ij}, \tag{4.15}$$

which is again equal to the GREG estimator of Y_i, given by

$$\hat{Y}_i^{GREG} = \hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)'\hat{\mathbf{B}}_i, \tag{4.16}$$

where $\hat{\mathbf{X}}_i = \sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}$ is the usual expansion estimator of $\mathbf{X}_i$ and $\hat{\mathbf{B}}_i = (\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}'_{ij})^{-1}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}y_{ij}$, provided that $\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}'_{ij}$ is non-singular.

A different calibration estimator of Y_i is obtained by minimizing the sum of distances at the population level subject to a constrain written also at the population level. In this case, the minimization problem is given by

$$\begin{aligned} \min \quad & \sum_{i=1}^{m}\sum_{j\in s_i}(h_{ij}-w_{ij})^2/w_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^{m}\sum_{j\in s_i} h_{ij}\mathbf{x}_{ij} = \mathbf{X}, \end{aligned} \tag{4.17}$$

assuming that the population total $\mathbf{X} = \sum_{i=1}^{m}\sum_{j=1}^{N_i}\mathbf{x}_{ij}$ is known. Solving (4.17) by Lagrange multipliers' method, where

$$L = \sum_{i=1}^{m}\sum_{j\in s_i}(h_{ij}-w_{ij})^2/w_{ij} + 2\boldsymbol{\lambda}'\left(\sum_{i=1}^{m}\sum_{j\in s_i} h_{ij}\mathbf{x}_{ij} - \mathbf{X}_i\right)$$

is the Lagrangian function and $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers, we obtain

$$\begin{aligned} h_{ij} &= w_{ij}(1+\mathbf{x}'_{ij}\boldsymbol{\lambda}),\ j \in s_i,\ i = 1,\ldots,m, \\ \boldsymbol{\lambda} &= \left(\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}'_{ij}\right)^{-1}(\mathbf{X} - \hat{\mathbf{X}}), \end{aligned} \tag{4.18}$$

see Appendix C.3. The resulting calibration estimator of the domain total Y_i is then

$$\hat{Y}_i^{LCALN} = \sum_{j \in s_i} h_{ij} y_{ij} = \hat{Y}_i + (\mathbf{X} - \hat{\mathbf{X}})' \hat{\mathbf{B}}_i^N, \tag{4.19}$$

where $\hat{\mathbf{B}}_i^N = \mathbf{T}^{-1} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} y_{ij}$, for $\mathbf{T} = \sum_{i=1}^{m} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \mathbf{x}'_{ij}$, provided that $\mathbf{T}$ is non-singular. Note that $\hat{Y}_i^{LCALN}$ is different from the GREG given in (4.16).

The advantage of the calibration estimator (4.15) (or GREG (4.16)) is that it reduces the bias since a different model is fitted for each area i. On the other hand, for those areas in which n_i is small, the variance of (4.15) may be large since only the area-specific data are used. The alternative calibration estimator given in (4.19) is expected to have larger bias since the calibration problem is solved at the national level. However, its variance will be smaller.

4.3 Small area estimation under cut-off sampling

We now study the case in which the sample drawn from each domain is obtained by cut-off sampling, that is, part of the domain U_i is excluded from sample selection. Under cut-off sampling, the domain U_i is partitioned into two strata, U_{iI} and U_{iE}. The stratum U_{iI} of size N_{iI} contains those units that can be potentially selected in the sample, called here the set of included (I) units. Stratum U_{iE} of size N_{iE} contains the excluded (E) units from the possible sample selection, with $N_i = N_{iI} + N_{iE}$. We want to estimate the domain total Y_i, which can be written as $Y_{iI} + Y_{iE}$, or the domain mean $\bar{Y}_i = Y_i/N_i$. The sample s_i of size n_i is drawn from the population of included individuals U_{iI}. Under this set up, the inclusion probabilities for the excluded units ($j \in U_{iE}$) are zero and the sampling weights for those units, do not exist. Thus, under cut-off sampling, it is not possible to obtain design-unbiased estimators. On the other hand, the use of naïve estimators obtained by ignoring the cut-off sampling, such as HT in (4.11) or HA in (4.12), might be severely design-biased. The design-bias of HA estimator is approximately given by

$$B_\pi(\hat{Y}_i^{HA}) \;=\; E_\pi(\hat{Y}_i) - Y_i \cong N_{iE}(\bar{Y}_{iI} - \bar{Y}_{iE}),$$

where here B_π and E_π denote bias and expectation under the sample replication mechanism respectively; $\bar{Y}_{iI} = Y_{iI}/N_{iI}$ and $\bar{Y}_{iE} = Y_{iE}/N_{iE}$ are the true means of the included and excluded units from area i respectively. This bias may be zero if there are not excluded units ($N_{iE} = 0$) or if $\bar{Y}_{iI} = \bar{Y}_{iE}$, which is not very common in the real cases where cut-off sampling is used, see Section 4.6. For HT estimator, the design-bias

is given by

$$B_\pi(\hat{Y}_i) \quad = \quad -Y_{iE},$$

which will be zero if the total for the excluded units is zero. Haziza et al. (2010) considered the use of calibration estimators to reduce the bias due to cut-off sampling. Precisely, they propose three types of calibration methods, namely direct calibration, calibration after reweighting and generalized calibration. In this section, we introduce these techniques and study their properties in the context of estimation in small domains or areas under cut-off sampling.

4.3.1 Direct calibration

Direct calibration estimators of domain totals are described in Section 4.2.2. The only difference here is that now the sample s_i comes from the set of included units U_{iI} only. The LCAL estimator is given by

$$\hat{Y}_i^{LCAL} = \hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)'\hat{\mathbf{B}}_i, \tag{4.20}$$

where $\hat{\mathbf{B}}_i = (\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}'_{ij})^{-1}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}y_{ij}$ is the weighted least square (WLS) estimator in the model

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \epsilon_{ij}, \tag{4.21}$$

where $E_m(\epsilon_{ij}) = 0$, $E_m(\epsilon_{ij}, \epsilon_{i\ell}) = 0$ for $j \neq \ell$ and $V_m(\epsilon_{ij}) = \sigma^2_\epsilon$. In order to study the properties of (4.20), we consider the theoretical estimator $\tilde{Y}_i^{LCAL}$ given by

$$\tilde{Y}_i^{LCAL} = Y_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)'\mathbf{B}_{iI}, \tag{4.22}$$

which is defined in terms of the census estimator $\mathbf{B}_{iI} = (\sum_{j\in U_{iI}} \mathbf{x}_{ij}\mathbf{x}'_{ij})^{-1}\sum_{j\in U_{iI}} \mathbf{x}_{ij}y_{ij}$ for the included units. Note that the sample s_i is drawn only from U_{iI} and thus $\hat{\mathbf{B}}_i$ estimates $\mathbf{B}_{iI}$. The design-bias of the calibration estimator in (4.22) is given by

$$B_\pi(\tilde{Y}_i^{LCAL}) = -(Y_{iE} - \mathbf{X}'_{iE}\mathbf{B}_{iI}), \tag{4.23}$$

see Appendix C.4. For the area mean, $\bar{Y}_i = Y_i/N_i$, the LCAL estimator becomes

$$\hat{\bar{Y}}_i^{LCAL} = \hat{Y}_i^{LCAL}/N_i. \tag{4.24}$$

As before, consider the theoretical estimator $\tilde{\bar{Y}}_i^{LCAL}$ given by

$$\tilde{\bar{Y}}_i^{LCAL} = \hat{\bar{Y}}_i + (\bar{\mathbf{X}}_i - \hat{\bar{\mathbf{X}}}_i)'\mathbf{B}_{iI}. \tag{4.25}$$

In Appendix C.4, we prove that the design bias of (4.25) is given by

$$B_\pi(\tilde{\bar{Y}}_i^{LCAL}) = \frac{N_{iE}}{N_i}[(\bar{Y}_{iI} - \bar{\mathbf{X}}'_{iI}\mathbf{B}_{iI}) - (\bar{Y}_{iE} - \bar{\mathbf{X}}'_{iE}\mathbf{B}_{iI})]. \tag{4.26}$$

This bias may be small if the proportion of excluded units is small. If the model holds for all the units in the population (included and excluded), then $E_m(\mathbf{B}_{iI}) = \boldsymbol{\beta}_i$, which is constant for the included and excluded units. In this case, the bias of $\tilde{\bar{Y}}_i^{LCAL}$ is approximately equal to the difference of means of the model errors in the included and excluded sets, which have expectation 0. In fact, taking expectation of (4.26) under the model (4.21), we obtain the bias under the model and the sampling replication mechanism, given by

$$\begin{aligned} B_{m,\pi}(\tilde{\bar{Y}}_i^{LCAL}) &= E_m\{B_\pi(\tilde{\bar{Y}}_i^{LCAL})\} \\ &= E_m\left\{\frac{N_{iE}}{N_i}[(\bar{Y}_{iI} - \bar{\mathbf{X}}'_{iI}\mathbf{B}_{iI}) - (\bar{Y}_{iE} - \bar{\mathbf{X}}'_{iE}\mathbf{B}_{iI})]\right\} \\ &= \frac{N_{iE}}{N_i}[(\bar{\mathbf{X}}'_{iI}\boldsymbol{\beta}_i - \bar{\mathbf{X}}'_{iI}\boldsymbol{\beta}_i) - (\bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i - \bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i)] \\ &= 0. \end{aligned} \tag{4.27}$$

In contrast, under the same model, the bias of the direct estimator (HA) is not zero unless the means of the auxiliary variables for the excluded and included units are equal:

$$\begin{aligned} B_{m,\pi}(\hat{\bar{Y}}_i^{HA}) &= E_m\{B_\pi(\hat{\bar{Y}}_i^{HA})\} \cong \frac{N_{iE}}{N_i}(\bar{Y}_{iI} - \bar{Y}_{iE}) \\ &= \frac{N_{iE}}{N_i}(\bar{\mathbf{X}}'_{iI}\boldsymbol{\beta}_i - \bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i) = \frac{N_{iE}}{N_i}(\bar{\mathbf{X}}_{iI} - \bar{\mathbf{X}}_{iE})'\boldsymbol{\beta}_i. \end{aligned} \tag{4.28}$$

The bias of the HT estimator is not zero unless the mean for the auxiliary variables of the excluded units is zero,

$$\begin{aligned} B_{m,\pi}(\hat{\bar{Y}}_i) &= E_m\{B_\pi(\hat{\bar{Y}}_i)\} = E_m\left\{-\frac{Y_{iE}}{N_i}\right\} \\ &= -\frac{\mathbf{X}'_{iE}\boldsymbol{\beta}}{N_i}. \end{aligned} \tag{4.29}$$

The problem is that there are no observations of the study variable for the excluded

units and, then, the model cannot be checked for those units. The design variance of $\tilde{Y}_i^{LCAL}$ and $\tilde{\bar{Y}}_i^{LCAL}$ is also given in Appendix C.4.

For the alternative calibration estimator (4.19), we consider again the theoretical version given by

$$\tilde{\bar{Y}}_i^{LCALN} = \hat{\bar{Y}}_i + (\bar{\mathbf{X}}_i - \hat{\bar{\mathbf{X}}}_i)'\mathbf{B}_{iI}^N, \tag{4.30}$$

where $\mathbf{B}_i^N = \mathbf{T}^{-1}\sum_{j\in U_{iI}}\mathbf{x}_{ij}y_{ij}$, for $\mathbf{T} = \sum_{i=1}^m\sum_{j\in U_{iI}}\mathbf{x}_{ij}\mathbf{x}_{ij}'$. The design bias of $\tilde{\bar{Y}}_i^{LCALN}$ is given by

$$B_\pi(\tilde{\bar{Y}}_i^{LCALN}) = \frac{N_{iE}}{N_i}\bar{Y}_{iI} - \frac{N_E}{N}\bar{\mathbf{X}}_I'\mathbf{B}_{iI}^N - \left(\frac{N_{iE}}{N_i}\bar{Y}_{iE} - \frac{N_E}{N}\bar{\mathbf{X}}_E'\mathbf{B}_{iI}^N\right),$$

see Appendix C.5. This bias will be small only if the model fitted at national level, holds for each area, with a common $\boldsymbol{\beta}$ for all areas.

4.3.2 Calibration after reweighting

This type of estimators consider that the allocation of each unit j into U_{iI} or U_{iE} is random, depending on a variable $c_{ij}|\mathbf{z}_{ij} \overset{iid}{\sim} \text{Bern}(p_{ij})$, where $\mathbf{z}_{ij}$ is a vector of auxiliary variables related to the probability of being in U_{iI}. We consider that the values $\mathbf{z}_{ij}$ are available for all the units in the population ($j \in U_i$). A common approach to obtain p_{ij} is to use logistic regression, that is, to consider that p_{ij} satisfies the model

$$p_{ij} = f(\mathbf{z}_{ij}, \boldsymbol{\zeta}) = \frac{\exp(\mathbf{z}_{ij}'\boldsymbol{\zeta})}{1+\exp(\mathbf{z}_{ij}'\boldsymbol{\zeta})},\ j = 1,\ldots,N_i,\ i = 1,\ldots,m, \tag{4.31}$$

where $\boldsymbol{\zeta}$ is the vector of regression coefficients. Then, p_{ij} can be estimated by $\hat{p}_{ij} = f(\mathbf{z}_{ij}, \hat{\boldsymbol{\zeta}})$, being $\hat{\boldsymbol{\zeta}}$ a consistent estimator of $\boldsymbol{\zeta}$ such as the maximum likelihood estimator (Haziza et al., 2010). Observe that this is the usual method to deal with non-response, since the two problems are mathematically equivalent.

Haziza et al. (2010) consider that the vectors of auxiliary variables $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are generally not the same, but $\mathbf{z}_{ij}$ should include variables that are related to both, the study variable, y_{ij}, and the probability of being included p_{ij}.

Analogously to the direct calibration estimator (4.20), calibration estimators after reweighting (RWCAL) can be obtained by finding a new set of weights $h_{ij}^* = w_{ij}^* F(\mathbf{x}_{ij}'\boldsymbol{\lambda}_i)$, where $w_{ij}^* = w_{ij}/\hat{p}_{ij}$, $j \in s_i$, such that the calibration equations

$$\sum_{j\in s_i} h_{ij}^*\mathbf{x}_{ij} = \mathbf{X}_i$$

are satisfied. The resulting estimator for the area total Y_i is given by

$$\hat{Y}_i^{RWCAL} = \sum_{j \in s_i} w_{ij}^* F(\mathbf{x}_{ij}' \boldsymbol{\lambda}_i) y_{ij}.$$

When $F(\mathbf{x}_{ij}' \boldsymbol{\lambda}_i) = 1 + \mathbf{x}_{ij}' \boldsymbol{\lambda}_i$, the RWCAL estimator of Y_i reduces to

$$\hat{Y}_i^* + (\mathbf{X}_i - \hat{\mathbf{X}}_i^*)' \hat{\mathbf{B}}_i^*, \tag{4.32}$$

where $\hat{Y}_i^* = \sum_{j \in s_i} w_{ij}^* y_{ij}$, $\hat{\mathbf{X}}_i^* = \sum_{j \in s_i} w_{ij}^* \mathbf{x}_{ij}$ and $\hat{\mathbf{B}}_{iI}^* = (\sum_{j \in s_i} w_{ij}^* \mathbf{x}_{ij} \mathbf{x}_{ij}')^{-1} \sum_{j \in s_i} w_{ij}^* \mathbf{x}_{ij} y_{ij}$. In order to obtain the design properties of the RWCAL estimator, again we define the corresponding theoretical estimator

$$\tilde{Y}_i^{RWCAL} = \tilde{Y}_i^* + (\mathbf{X}_i - \tilde{\mathbf{X}}_i^*)' \mathbf{B}_i, \tag{4.33}$$

where $\tilde{Y}_i^* = \sum_{j \in s_i} w_{ij} y_{ij} / p_{ij}$, $\tilde{\mathbf{X}}_i^* = \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} / p_{ij}$. It is easy to see that, when the model (4.31) holds even though (4.21) does not hold, $\tilde{Y}_i^{RWCAL}$ is unbiased under the sample design and the included selection mechanism, that is,

$$B_{\pi,p}(\tilde{Y}_i^{RWCAL}) = E_p E_{\pi|p}(\tilde{Y}_i^{RWCAL}) - Y_i = 0,$$

where now the expectations are taken under the sample replication mechanism and included selection mechanism respectively, see Appendix C.6. If the model (4.21) holds, the estimator $\tilde{Y}_i^{RWCAL}$ is asymptotically unbiased under the model and the sampling design, that is

$$B_{m,\pi}(\tilde{Y}_i^{RWCAL}) = E_m E_{\pi|m}(\tilde{Y}_i^{RWCAL}) - Y_i = 0,$$

where $B_{m,\pi}$, denotes the bias and under the sample replication mechanism and the model.

For the area mean $\bar{Y}_i$, the RWCAL estimator is given by

$$\hat{\bar{Y}}_i^{RWCAL} = \hat{\bar{Y}}_i^* + (\bar{\mathbf{X}}_i - \hat{\bar{\mathbf{X}}}_i^*) \hat{\mathbf{B}}_i^* \tag{4.34}$$

where $\hat{\bar{Y}}_i^* = \sum_{j \in s_i} w_{ij}^* y_{ij} / N_i$ and $\hat{\bar{\mathbf{X}}}_i = \sum_{j \in s_i} w_{ij}^* \mathbf{x}_{ij} / N_i$.

4.3.3 Generalized calibration estimators

Generalized calibration estimators are described in Kott (2006) and Deville et al. (1993) among others. The idea of generalized calibration is to remove the requirement of minimizing a distance function. Instead, we obtain a new set of weights of the form

$\tilde{h}_{ij} = w_{ij}F(\boldsymbol{\lambda}_i'\mathbf{z}_{ij})$, which satisfy the calibration equation

$$\sum_{j\in s_i} \tilde{h}_{ij}\mathbf{x}_{ij} = \mathbf{X}_i.$$

For $F(\boldsymbol{\lambda}_i'\mathbf{z}_{ij}) = 1 + \boldsymbol{\lambda}_i'\mathbf{z}_{ij}$, the new weights $\tilde{h}_{ij}$ are given by

$$\tilde{h}_{ij} = w_{ij}\left[1 + (\mathbf{X}_i - \hat{\mathbf{X}}_{iI})'\left(\sum_{j\in s_i} w_{ij}\mathbf{z}_{ij}\mathbf{x}_{ij}'\right)^{-1}\mathbf{z}_{ij}\right], \tag{4.35}$$

see Appendix C.7. The advantage of generalized calibration with respect to calibration after reweighting is that the vector $\mathbf{z}_{ij}$ of auxiliary variables is required only for $j \in s_i$. This vector might include the target variable y_{ij}, as one component, which may help to reduce the bias (Haziza et al., 2010). The resulting generalized calibration (GCAL) estimator for the domain total Y_i is

$$\hat{Y}_i^{GCAL} = \sum_{j\in s_i} w_{ij}F(\boldsymbol{\lambda}_i'\mathbf{z}_{ij})y_{ij}.$$

When the functional form of the new weights is given by $F(\boldsymbol{\lambda}_i'\mathbf{z}_{ij}) = 1+\boldsymbol{\lambda}_i'\mathbf{z}_{ij}$, the GCAL estimator can be expressed as

$$\hat{Y}_i^{GCAL} = \hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)'\hat{\tilde{\mathbf{B}}}_i, \tag{4.36}$$

where $\hat{\tilde{\mathbf{B}}}_i = \hat{\tilde{T}}_i^{-1}\hat{\tilde{t}}_i$, $\hat{\tilde{t}}_i = \sum_{j\in s_i} w_{ij}\mathbf{z}_{ij}y_{ij}$ and $\hat{\tilde{T}}_i = \sum_{j\in s_i} w_{ij}\mathbf{z}_{ij}\mathbf{x}_{ij}'$. $\hat{\tilde{\mathbf{B}}}_i$ can be seen as the estimated regression coefficient fitted by the procedure of instrumental variables, with $\mathbf{z}_{ij}$ acting as the vector of instrumental variables. If $\mathbf{z}_{ij}$ explains well the allocation into included and excluded units, the bias of (4.36) is small. Haziza et al. (2010) show that the bias under the sampling design and the included/excluded selection mechanism of $\hat{Y}_i^{GCAL}$ is given by

$$B_{p,\pi}(\hat{Y}_i^{GCAL}) = E_pE_{\pi|p}(\hat{Y}_i^{GCAL}) - Y_i \approx -\sum_{j\in U_i}(1-p_{ij})(y_{ij} - \mathbf{x}_{ij}'\tilde{\mathbf{B}}_p),$$

where $\tilde{\mathbf{B}}_p = (\sum_{j\in U_i} p_{ij}\mathbf{z}_{ij}\mathbf{x}_{ij}')^{-1}\sum_{j\in U_i} p_{ij}\mathbf{z}_{ij}y_{ij}$. If $\mathbf{z}_{ij}$ explains the allocation into included and excluded individuals, the estimator (4.36) has a small bias. For the area mean $\bar{Y}_i$, the GCAL estimator is given

$$\hat{\bar{Y}}_i^{GCAL} = \hat{\bar{Y}}_i + (\bar{\mathbf{X}}_i - \hat{\bar{\mathbf{X}}}_i)\hat{\tilde{\mathbf{B}}}_i. \tag{4.37}$$

4.3.4 EBLUP

Estimators described so far use mainly the information coming from the domain. When the sample size n_i is small, these estimators are inefficient. Small area estimation methods, or indirect estimators, reduce the variance by increasing the effective sample size. The book by Rao and Molina (2015) contains a comprehensive account of small area estimation methods. In this subsection, we focus on the methods that are based on models (model-based) rather than assisted by models. This means that the estimators have good properties under the distribution induced by the model. However, we will also study their properties under the design.

The models can be stated at the domain or at the unit level. Area level models relate the direct estimators for the domains to domain-specific covariates. Unit level models relate the unit values of a study variable to unit specific covariates. The basic area or unit-level models are particular cases of linear mixed models.

Here we consider a very popular unit level model, which is the nested error model introduced by Battese et al. (1988). This model establishes a linear relationship between the target variable for the population units y_{ij} and the auxiliary variables $\mathbf{x}_{ij}$, for all the areas, as follows

$$\begin{aligned} &y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij},\ v_i \overset{iid}{\sim} N(0, \sigma_v^2), \\ &e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2),\ j = 1, \ldots, N_i,\ i = 1, \ldots, m, \end{aligned} \tag{4.38}$$

where v_i is the effect of area i and e_{ij} is the residual error. Area effects v_i and errors e_{ij} are assumed to be mutually independent. We denote by $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_v^2, \sigma_e^2)'$ the vector of unknown parameters in the population model (4.38).

Let us write the model (4.38) in matrix notation. For this, we define the area vectors and matrices

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{iN_i})',\ \mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i})',\ \mathbf{e}_i = (e_{i1}, \ldots, e_{iN_i})',\ i = 1, \ldots, m.$$

Then, in matrix notation, the model reads

$$\mathbf{y}_i \overset{ind}{\sim} N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i),\ \mathbf{V}_i = \sigma_v^2 \mathbf{1}_{N_i}\mathbf{1}_{N_i}' + \sigma_e^2 \mathbf{I}_{N_i},\ i = 1, \ldots, m, \tag{4.39}$$

where $\mathbf{1}_k$ denotes a vector of ones of size k and $\mathbf{I}_k$ is the $k \times k$ identity matrix. Consider also the population vectors and matrices

$$\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_m')',\ \mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_m')',\ \mathbf{e} = (\mathbf{e}_1', \ldots, \mathbf{e}_m')'.$$

Additionally, define the block-diagonal matrix $\mathbf{Z} = \text{diag}_{1\leq i\leq m}(\mathbf{1}_{N_i})$ and the vector $\mathbf{v} = (v_1, \ldots, v_m)'$. Then, the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e},\ \mathbf{v} \sim N(\mathbf{0}_m, \sigma_v^2\mathbf{I}_m),\ \mathbf{e} \sim N(\mathbf{0}_N, \sigma_e^2\mathbf{I}_N),$$

or, equivalently,

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}),\ \mathbf{V} = \sigma_v^2\mathbf{Z}\mathbf{Z}' + \sigma_e^2\mathbf{I}_N = \text{diag}_{1\leq i\leq m}(\mathbf{V}_i).$$

Let $H = \mathbf{b}'\mathbf{y}$ be a linear parameter, where $\mathbf{b}$ is a non-stochastic vector. Let us decompose into sample (drawn from included units in each area U_{iI}) and out-of-sample elements the population vector $\mathbf{y}$, the design matrix $\mathbf{X}$ and the covariance matrix $\mathbf{V}$,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix},\quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix},\quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_r \end{pmatrix}.$$

The linear parameter H can be decomposed accordingly as $H = \mathbf{b}_s'\mathbf{y}_s + \mathbf{b}_r'\mathbf{y}_r$. Under the model (4.38), the best linear unbiased predictor (BLUP) of H is the linear function of the sample data $\tilde{H} = \boldsymbol{\alpha}_s'\mathbf{y}_s$ which is model unbiased and minimizes the model mean squared error (MSE), given by $\text{MSE}_m(\tilde{H}) = E_m(\tilde{H} - H)^2$. The BLUP is given by

$$\tilde{H}^{BLUP} = \mathbf{b}_s'\mathbf{y}_s + \mathbf{b}_r'[\mathbf{X}_r\tilde{\boldsymbol{\beta}}_s + \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s'\tilde{\boldsymbol{\beta}}_s)], \tag{4.40}$$

where $\tilde{\boldsymbol{\beta}}_s$ is the weighted least squared estimator of $\boldsymbol{\beta}$ given by

$$\tilde{\boldsymbol{\beta}}_s = \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s.$$

The BLUP, $\tilde{H}^{BLUP}$, depends on $\boldsymbol{\theta}$, which is typically unknown. Replacing $\boldsymbol{\theta}$ by a consistent estimator $\hat{\boldsymbol{\theta}}$, we obtain the so called empirical BLUP (EBLUP), and denoted $\hat{H}^{EBLUP} = \tilde{H}^{BLUP}(\hat{\boldsymbol{\theta}})$.

When estimating the domain mean $H = \bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$, the vector $\mathbf{b}$ is given by $\mathbf{b} = (\mathbf{0}_{N_1}', \ldots, \mathbf{0}_{N_{i-1}}', N_i^{-1}\mathbf{1}_{N_i}', \mathbf{0}_{N_i+1}', \ldots, \mathbf{0}_{N_m}')'$, where $\mathbf{0}_k$ denotes a vector of zeros of size k. If the domain sampling fraction, n_i/N_i, is negligible, the BLUP estimator of $\bar{Y}_i$ is a weighted average of the GREG estimator for the domain mean under simple random sampling and the regression synthetic estimator $\bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_s$ (Rao and Molina, 2015), that is,

$$\hat{\bar{Y}}_i^{BLUP} \cong \gamma_{is}[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\hat{\boldsymbol{\beta}}_s] + (1 - \gamma_{is})\bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_s, \tag{4.41}$$

where $\gamma_{is} = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_i)$. Thus, for domains with large sample size n_i, $\hat{\bar{Y}}_i^{BLUP}$ approaches the GREG estimator $\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\hat{\boldsymbol{\beta}}_s$, whereas for domains with small

sample size n_i, $\hat{\bar{Y}}_i$ borrows strength from the other domains by approaching the regression synthetic estimator $\bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_s$.

Again, in order to study the properties of $\hat{\bar{Y}}_i^{BLUP}$, we consider the theoretical estimator using the true value of $\boldsymbol{\beta}$, assuming that model (4.38) holds for all the population units,

$$\tilde{\bar{Y}}_i^{BLUP} = \gamma_{is}[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\boldsymbol{\beta}] + (1-\gamma_{is})\bar{\mathbf{X}}_i'\boldsymbol{\beta}.$$

The design-bias of $\tilde{\bar{Y}}_i^{BLUP}$ is given by

$$B_\pi(\tilde{\bar{Y}}_i^{BLUP}) = \gamma_{is}\frac{N_{iE}}{N_{iI}}\left[(\bar{Y}_i - \bar{\mathbf{X}}_i'\boldsymbol{\beta}) - (\bar{Y}_{iE} - \bar{\mathbf{X}}_{iE}'\boldsymbol{\beta})\right] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i),$$

see Appendix C.8. This bias will be small if the same model (4.38) holds for the whole population and if the ratio of excluded over included individuals is small. In fact,

$$\begin{aligned} & B_{m,\pi}(\tilde{\bar{Y}}_i^{BLUP}) \\ = \; & E_m\left\{\gamma_{is}\frac{N_{iE}}{N_{iI}}\left[(\bar{Y}_i - \bar{\mathbf{X}}_i'\boldsymbol{\beta}) - (\bar{Y}_{iE} - \bar{\mathbf{X}}_{iE}'\boldsymbol{\beta})\right] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i)\right\} \\ = \; & \gamma_{is}\frac{N_{iE}}{N_{iI}}\left[(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{\mathbf{X}}_i'\boldsymbol{\beta}) - (\bar{\mathbf{X}}_{iE}'\boldsymbol{\beta} - \bar{\mathbf{X}}_{iE}'\boldsymbol{\beta})\right] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{\mathbf{X}}_i'\boldsymbol{\beta}) \\ = \; & 0. \end{aligned} \tag{4.42}$$

The design-variance of $\tilde{\bar{Y}}_i^{BLUP}$ is given in Appendix C.8.

4.3.5 Empirical best predictor

For estimation of more complex non-linear parameters, the BLUP has no meaning and we need to resort the methods dealing with more general parameters such as the best/Bayes predictor of Molina and Rao (2010). Special non-linear parameters are poverty and inequality indicators which are functions of welfare/expenses. The best predictor can also be used for estimation of characteristics such as median, quantiles or even the empirical distribution function of the variable of interest. Additionally, it can be used for estimation of totals and means of a target variable, when the dependent variable in the model is a transformation (e.g. log or square root) of the target variable, which occurs in cases of non normality or heteroscedasticity. For a parameter that is a linear function of the values of the dependent variable in a linear model, the best predictor equals to the BLUP.

Let $y_{T,ij}$ be a one-to-one transformation of the target variable y_{ij}, that is, $y_{T,ij} =$

$T(y_{ij})$. Consider that $y_{T,ij}$ follows the nested error model

$$y_{T,ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + e_{ij},\ v_i \overset{iid}{\sim} N(0, \sigma_v^2),$$
$$e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2),\ j = 1, \ldots, N_i,\ i = 1, \ldots, m, \tag{4.43}$$

Defining $\mathbf{y}_{T,i} = (y_{T,i1}, \ldots, y_{T,iN_i})'$. the model is

$$\mathbf{y}_{T,i} \overset{ind}{\sim} N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i),\ \mathbf{V}_i = \sigma_v^2\mathbf{1}_{N_i}\mathbf{1}'_{N_i} + \sigma_e^2\mathbf{I}_{N_i},\ i = 1, \ldots, m. \tag{4.44}$$

Let us decompose into sample and out-of-sample elements the domain vector $\mathbf{y}_{T,i} = ((\mathbf{y}_{T,is})', (\mathbf{y}_{T,ir})')'$. Define a general parameter $H_i = h(\mathbf{y}_{T,i})$ as a function of the domain population vector $\mathbf{y}_{T,i}$. The best predictor of $H_i = h(\mathbf{y}_{T,i})$ is defined as the function of the sample observations of the domain $\mathbf{y}_{T,is}$ that minimizes the model MSE, and is given by

$$\tilde{H}_i^B(\boldsymbol{\theta}) = E_{\mathbf{y}_{T,ir}}[H_i|\mathbf{y}_{T,is}; \boldsymbol{\theta}], \tag{4.45}$$

where the expectation is taken with respect to the distribution of $\mathbf{y}_{T,ir}|\mathbf{y}_{T,is}$, which depends on the true value of $\boldsymbol{\theta}$. $\tilde{H}_i^B(\boldsymbol{\theta})$ is unbiased with respect to the model regardless of the complexity of the function $h(\cdot)$. However, it cannot be calculated in practice since model parameters $\boldsymbol{\theta}$ are typically unknown. An empirical best (EB) predictor of H_i, denoted as $\hat{H}_i^{EB}$, is then obtained by replacing $\boldsymbol{\theta}$ in $\tilde{H}_i^B(\boldsymbol{\theta})$ by a consistent estimator $\hat{\boldsymbol{\theta}}$, that is, $\hat{H}_i^{EB} = \tilde{H}_i^B(\hat{\boldsymbol{\theta}})$. The EB is not exactly unbiased, but the bias arising from the estimation of $\boldsymbol{\theta}$ is typically negligible when the overall sample size n is large.

For some non-linear parameters, the expectation given in (4.45) cannot be calculated analytically; in those cases, $\hat{H}_i^{EB}$ can be approximated by Monte Carlo as proposed in Molina and Rao (2010). This is done by simulating L replicates $y_{T,ij}^{(\ell)}; \ell = 1, \ldots, L$ of $y_{T,ij}$, $j \in r_i$, where r_i are the non-sample units of area i, attaching the sample elements $y_{T,ij}$, $j \in s_i$ to form the domain population vector $\mathbf{y}_{T,i}^{(\ell)}$, calculating the corresponding $H_i^{(\ell)} = h(\mathbf{y}_{T,i}^{(\ell)})$ for each ℓ and averaging over the L replicates as $\hat{H}_i^{EB} = L^{-1}\sum_{\ell=1}^{L} H_i^{(\ell)}$. For further details, see Molina and Rao (2010).

4.4 MSE estimation

The EB of Section 4.3.5 and the GREG under cut-off sampling in Section 4.2.2 are all based on a linear regression model. Moreover, the design MSE is unstable in domains with small sample size. Thus, here we will estimate the model MSEs of these estimators. Our model MSE estimators are obtained using the same bootstrap

procedure of Molina and Rao (2010), which is based on the parametric bootstrap method for finite populations of González-Manteiga et al. (2008). According to this procedure, the bootstrap MSE of $\hat{H}_i^{EB}$ under the nested error model (4.38) is obtained as follows: i) Fit the model (4.38) to the sample data drawn from the population of included units, $(\mathbf{y}_{T,s}, \mathbf{X}_s)$, obtaining estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ of $\boldsymbol{\beta}$, σ_v^2 and σ_e^2 respectively. ii) For $b = 1, \ldots, B$, with B large, generate independently $v_i^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_v^2)$ and $e_{ij}^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. iii) Construct B iid bootstrap population vectors $\mathbf{y}_{T,i}^{*(b)}$ with elements $y_{T,ij}^{*(b)}$ generated as

$$y_{T,ij}^{*(b)} = \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} + v_i^{*(b)} + e_{ij}^{*(b)}, \ b = 1 \ldots, B.$$

From each bootstrap population b, calculate the general parameter $H_i^{*(b)} = h(\mathbf{y}_{T,i}^{*(b)})$ for $b = 1, \ldots, B$. iv) From each bootstrap population b, take the sample with the same indices as the initial sample s and, using the sample elements $\mathbf{y}_{T,is}^{*(b)}$ and the known population vectors $\mathbf{x}_{ij}$, $j \in U_i$, calculate the bootstrap EB predictor of H_i, denoted by $\hat{H}_i^{EB*(b)}$, $b = 1, \ldots, B$. v) A bootstrap estimator of $\text{MSE}_m(\hat{H}_i^{EB})$ is then

$$\text{mse}_m(\tilde{H}_i^{EB}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{H}_i^{EB*(b)} - H_i^{*(b)})^2. \tag{4.46}$$

A bootstrap estimator of $\text{MSE}_m(\hat{Y}_i^{GREG})$ can be obtained using the same procedure. The only difference is that the error ϵ_{ij} in the GREG model (4.21) is assumed to be the sum of the area effect v_i and the model error e_{ij} in model (4.38). If H_i is a linear parameter and there is no transformation of the target variable, that is, $H_i = \mathbf{b}_i'\mathbf{y}_i$ and $T(y_i) = y_i$, the EB predictor equals the EBLUP. In this case, (4.46) gives an estimator of $\text{MSE}_m(\hat{H}_i^{EBLUP})$. This naïve bootstrap estimator is first-order unbiased in the sense that its model bias is $O(m^{-1})$, but not $o(m^{-1})$. There are bias corrections, but those corrections increase the variance and may yield negative MSE estimates. In the literature we cannot find bootstrap estimators of the MSE that are strictly positive and also second-order unbiased.

4.5 Simulation experiment

This section compares the performance of direct calibration and small area estimation methods when the sample is drawn by cut-off sampling. Specifically, we will compare the two calibration estimators proposed in Section 4.3.1, LCAL and LCALN, the naïve HT direct estimator (4.11) ignoring the cut-off sampling and the EBLUP of the domain mean $\bar{Y}_i$.

Calibration estimators preserve good properties under the design even when the model does not hold. Since under the model, the EBLUP of a linear parameter is known to be approximately the most efficient linear and unbiased estimator, here we want to compare its design-based properties with those of the calibration estimators. For this reason, we run design-based simulations by generating one population vector $\mathbf{y}$, keeping it fixed and repeatedly drawing samples from it. The population vector $\mathbf{y}$ is generated from the nested error model in (4.38). To allocate the units into the set of included and excluded units, we generate a binary variable c_{ij} for each $j = 1, \ldots, U_i$ and $i = 1, \ldots, m$, where $c_{ij} = 0$ if $j \in U_{iE}$ and $c_{ij} = 1$ otherwise. In each Monte Carlo (MC) replicate, we draw a srswor from those units with $c_{ij} = 1$ independently for each domain i, $i = 1, \ldots, m$.

The simulations were implemented in the statistical software environment R (R development core team 2016) using the packages `sampling` (Tillé and Matei, 2016), `nlme` (Pinheiro et al., 2017) and `sae` (Molina and Marhuenda, 2015). The first package contains functions for drawing samples and obtaining calibration estimators. The `nmle` package fits Gaussian linear and nonlinear mixed-effects models. The `sae` package contains functions for small area estimation.

We consider a population of $N = 20,000$ individuals divided into $m = 80$ domains with the same size $N_i = 250$, $i = 1, \ldots, m$. We generate values x_{ijq} for three auxiliary variables $q = 1, 2, 3$, each generated from a $N(3, 2)$. The variables c_{ij} are generated independently for each j and i from a Bernoulli distribution with probability $p_{ij} = \Pr(c_{ij} = 1)$, which is related to the auxiliary variables $x_{q,ij}$ through a logit model, that is,

$$p_{ij} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\zeta})}{1 + \exp(\mathbf{x}'_{ij}\boldsymbol{\zeta})}, \; j = 1, \ldots, N_i, \; i = 1, \ldots, m.$$

We choose $\boldsymbol{\zeta} = (0.75, 1, 1)'$. With these model parameters, the set of included units, that is, those with $c_{ij} = 1$, represent approximately half of the total population.

We generate the values of the target variable y_{ij} from those of the auxiliary variables $(x_{1,ij}, x_{2,ij}, x_{3,ij})'$, such that the coefficient of determination is approximately 0.5. To achieve that, the vector of regression coefficients is taken as $\boldsymbol{\beta} = (1, 1.5, 1)'$, the domain effects standard deviation (sd) and error sd are respectively taken as $\sigma_u = 0.75$ and $\sigma_e = 4$.

We draw $L = 1,000$ samples from those units with $c_{ij} = 1$, $j \in U_i$ by independent simple random sampling without replacement (srswor) of size n_i for each domain $i = 1, \ldots, m$, taking $n_i = 5$, $1 \leq i \leq 20$, $n_i = 10$, $21 \leq i \leq 40$, $n_i = 30$, $41 \leq i \leq 60$ and $n_i = 50$, $61 \leq i \leq 80$.

With the sample data from the ℓ-th MC replicate, we compute the direct HT estimator in (4.11), the calibration estimators with calibration at the domain level (LCAL) and at the population level (LCALN) and the EBLUP of $\bar{Y}_i$. We do not report results for the calibration after reweighting (RWCAL) and the generalized calibration (GCAL) estimator, because in our simulations they showed instability. To obtain the new weights, h_{ij}, in the calibration estimators, we use the function *calib* from package sampling (Tillé and Matei, 2016). The EBLUP estimators are computed using the REML method for estimation of the model parameters σ_v^2, σ_e^2 and $\boldsymbol{\beta}$.

Let $\hat{\bar{Y}}_i$ be a generic estimator (HT, LCAL, LCALN or EBLUP) of $\bar{Y}_i$ and $\hat{\bar{Y}}_i^{(b)}$ its value obtained in MC replicate b. We evaluate the performance of estimators in terms of relative bias (RB) and relative root MSE (RRMSE) under the design, approximated empirically as

$$\mathrm{RB}_\pi(\hat{\bar{Y}}_i) = 100\frac{B^{-1}\sum_{b=1}^{B}(\hat{\bar{Y}}_i^{(b)} - \bar{Y}_i)}{\bar{Y}_i}, \quad \mathrm{RRMSE}_\pi(\hat{\bar{Y}}_i) = 100\frac{\sqrt{B^{-1}\sum_{b=1}^{B}(\hat{\bar{Y}}_i^{(b)} - \bar{Y}_i)^2}}{\bar{Y}_i}.$$

Averages across domains of absolute RB and of RRMSE are also calculated as

$$\overline{\mathrm{ARB}} = m^{-1}\sum_{i=1}^{m}|\mathrm{RB}_\pi(\hat{\bar{Y}}_i)|, \quad \overline{\mathrm{RRMSE}} = m^{-1}\sum_{i=1}^{m}\mathrm{RRMSE}_\pi(\hat{\bar{Y}}_i).$$

Figure 4.1: Percent RB (left) and RRMSE (right) of HT, LCAL, LCALN and EBLUP estimators of domain mean, $\bar{Y}_i$, for each area.

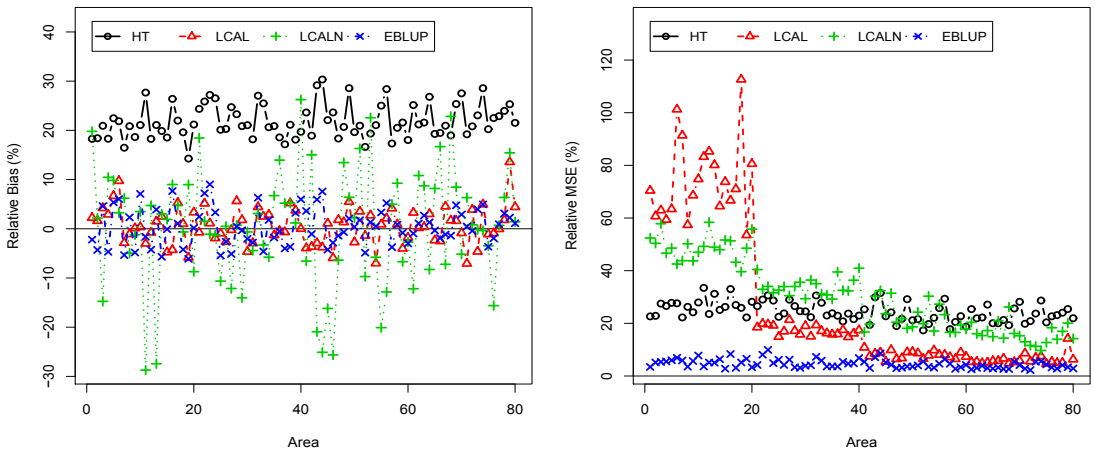


Figure 4.1 displays the percent RB (left) and RRMSE (right) for the considered

estimators of the mean $\bar{Y}_i$ for each area i (x-axis), under srswor within the included elements ($c_{ij} = 1$) in each area i. This figure shows that the HT direct estimator obtained ignoring the cut-off sampling has large design bias and MSE for all the areas. On the other hand, the LCALN estimator shows a large bias for several areas, which may be due to the fact of not taking into account the area effect, since the minimization is done at the national level and the restriction is also established at the national level. LCAL estimator is the best in terms of bias. Note that this estimator fits a different regression parameter $\boldsymbol{\beta}_i$ for each area. Moreover, Figure 4.1 right shows very large RRMSEs for the two calibration estimators, LCAL and LCALN, for those areas with the smallest sample sizes ($n_i \leq 20$). See that EBLUP exhibits the best results in terms of MSE and keeps a small design bias. In fact, the difference between the EBLUP and LCAL estimators in terms of bias is small.

Table 4.1: Averages across areas of percent absolute RB and RRMSE and average B_π^2/MSE_π for HT, LCAL, LCALN and EBLUP (in percentage).

Method	$\overline{\text{ARB}}$	$\overline{\text{RRMSE}}$	B_π^2/MSE_π
HT	21.82	24.45	98.32
LCAL	2.96	27.33	2.48
LCALN	8.97	30.44	0.04
EBLUP	3.13	4.56	0.18

Table 4.1 displays the $\overline{\text{ARB}}$, the $\overline{\text{RRMSE}}$ and the squared bias over the MSE under the design (in percentage) for the considered estimators. In this table, again HT exhibits large design-bias and its B_π^2/MSE_π ratio is practically 100%, whereas the considered calibration estimators and the EBLUP reduce considerably the bias. Again, EBLUP shows the best performance in terms of efficiency. The LCALN estimator performs the best in terms of ratio B_π^2/MSE_π because it has a large MSE, so we consider that LCAL estimator performs better, although EBLUP is clearly performing the best when considering both MSE and bias under the design.

4.6 Estimation of sales of tobacco products

In this section, we compare the performance of calibration and EB estimators of the total sales of a particular tobacco product in the Spanish provinces. We use data from an important tobacco distribution company in Spain. This data set contains the purchases of that particular product made by the tobacco shops in all the provinces. Note that in Spain, tobacco is sold only by authorized shops. Our target domains are $m = 48$ Spanish provinces. Canary Islands, Ceuta and Melilla are not included in the data set. This data set contains the volume of sales (y_{ij}, in Euros) of the considered product in

November 2016 of a selection of the tobacco shops and the volume of purchases (z_{ij} in Euros) during the previous three months for practically all the tobacco shops (units) in the mentioned provinces. It also contains a variable indicating whether the tobacco shop is supplied with a device registering the necessary information about each sale. This device is set in those tobacco shops with larger sales. These are the tobacco shops for which the volume of sales of the considered product is available. These shops j with both z_{ij} and y_{ij} available for a province i compose the set of included units U_{iI}, which equals the sample s_i in this case. In Appendix C.9, we prove that with $s_i = U_{iI}$, we obtain the same direct, calibration and EBLUP estimators as when the sample s_i (U_{iI}) is considered to be drawn from U_i by srswor.

The total sample size (shops with the device registering the sales) is $n = 1,842$ and the population size (total number of tobacco shops) is $N = 12,791$. We will consider the nested error model given in (4.43). As auxiliary variable, we consider the purchases of tobacco shops (z_{ij}). In a preliminary study, we have seen that both, purchases and sales, show skewness to the right. Moreover, when fitting a linear model for the sales in terms of purchases, the resulting residuals display a mild pattern of heteroscedasticity. Transforming the sales (y_{ij}) and the purchases (z_{ij}) with the squared root, that is, taking $y_{T,ij} = y_{ij}^{1/2}$ and $\mathbf{x}_{ij} = (1, x_{ij})'$, with $x_{ij} = z_{ij}^{1/2}$ in the nested error model, seems to solve the problem. Then the EB predictors of the total sales for province i, $Y_i = \sum_{j=1}^{N_i} y_{ij} = \sum_{j=1}^{N_i} y_{T,ij}^2 = h(\mathbf{y}_{T,i})$ will be given by

$$\hat{Y}_i^{EB} = E_{\mathbf{y}_{T,ir}}[h(\mathbf{y}_{T,i})|\mathbf{y}_{T,is}; \boldsymbol{\theta}].$$

We wish to compare the EBs and the calibration estimators of the province totals Y_i in terms of coefficients of variations (CVs) or estimated RRMSEs. We compute calibration estimates by minimizing the chi-squared distance in the province subject to the calibration for that province. Thus, we are considering the LCAL or GREG estimator given in (4.16). We also compare with the basic direct (DIR) estimator of the total sales, which, in this case, where $s_i = U_{iI}$, reduces to

$$\hat{Y}_i = \sum_{j \in U_{iI}} y_{ij}.$$

In absence of cut-off sampling and under srswor, the MSE of the above direct estimator equals its variance. A design-unbiased estimator of the design MSE (or variance) is given by

$$\text{mse}_\pi(\hat{Y}_i) = N_i^2 \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right),$$

where $s_i^2 = (n_i - 1)^{-1} \sum_{j \in s_i} (y_{ij} - \bar{y}_{is})^2$ is the sample variance. Note that, in our case,

the srswor is applied after cutting-off the excluded population and the bias due to this is not accounted for in the above MSE estimator. For the EB predictor, the MSE was estimated using the parametric bootstrap described in Section 4.4, taking $H^{*(b)} = Y_i^{*(b)}$ and $\hat{H}^{EB*(b)} = \hat{Y}_i^{EB*(b)}$. The MSE for the GREG estimator is estimated using the same procedure, taking instead $\hat{H}^{EB*(b)} = \hat{Y}_i^{GREG*(b)}$. However, the GREG estimator is not defined for non-linear parameters such as $Y_i = h(\mathbf{y}_{T,i})$ under the square root transformation of y_{ij}. Thus, here we calculate the GREG based on a linear model like (4.21) for y_{ij} in terms of $\mathbf{x}_{ij}$ without transformation of y_{ij}. The resulting bootstrap estimator of the model MSE of the GREG estimator includes the error due to the fact that the model assumed by the GREG estimator is not correct. Before comparing these estimates, let us analyze the goodness of fit of the considered model (4.43). First, we look at the residuals obtained from the model fit, $\hat{e}_{ij} = y_{T,ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} - \hat{v}_i$, against predicted values $\hat{y}_{T,ij} = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}} + \hat{v}_i$ in Figure 4.2 (left) and the histogram of residuals on the right plot. Figure 4.2 left shows few negative outliers, which agrees with a slightly larger left tail in the histogram (right). Apart from that, residuals do not exhibit any remarkable pattern. In fact, residuals appear to be very much concentrated around zero, which indicates a high predictive power of the model.

Figure 4.2: EB residuals against predicted values (left), and histogram of EB residuals (right).

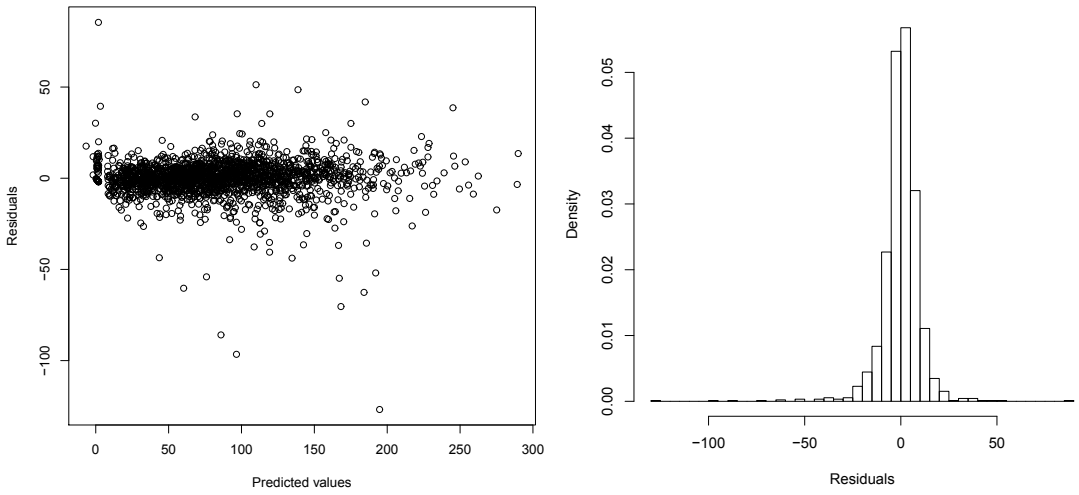
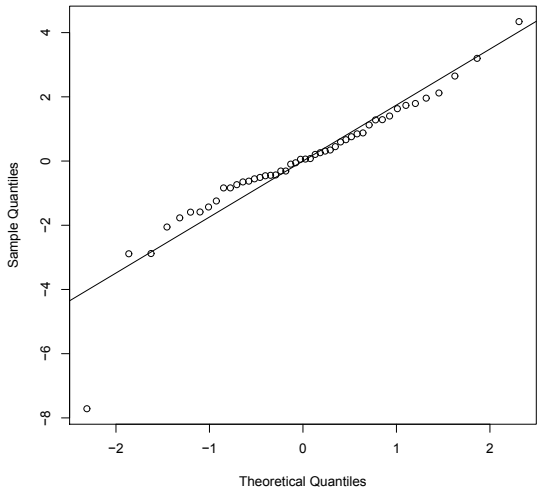


Figure 4.3 shows the Normal Q-Q plot of predicted area effects $\hat{v}_i$. This plot supports the normality of v_i except for one outlier appearing at the left tail of the distribution. This point corresponds to the province with the smallest sample size ($n_i = 3$ observations). This means that, for that province, the estimated random effect $\hat{v}_i$ is not very reliable.

Figure 4.4 left shows EB predictors against direct estimates of the total sales of the

Figure 4.3: Normal Q-Q plot of predicted province effects $\hat{v}_i$.



considered tobacco product for each Spanish province. We indicate the province sample sizes in the point labels. This table shows that all points corresponding to provinces with small sample sizes (lower left corner) are either on the line or very close to it, that is, the two estimates are very similar. However, for the two provinces with the largest sample sizes, the EB estimates are slightly larger than the corresponding direct estimates. Figure 4.4 right displays EB against GREG estimates. In this plot, all points are basically on the equality line.

Figure 4.4: EB estimates against direct estimates (left), and against GREG estimates (right) for each province.

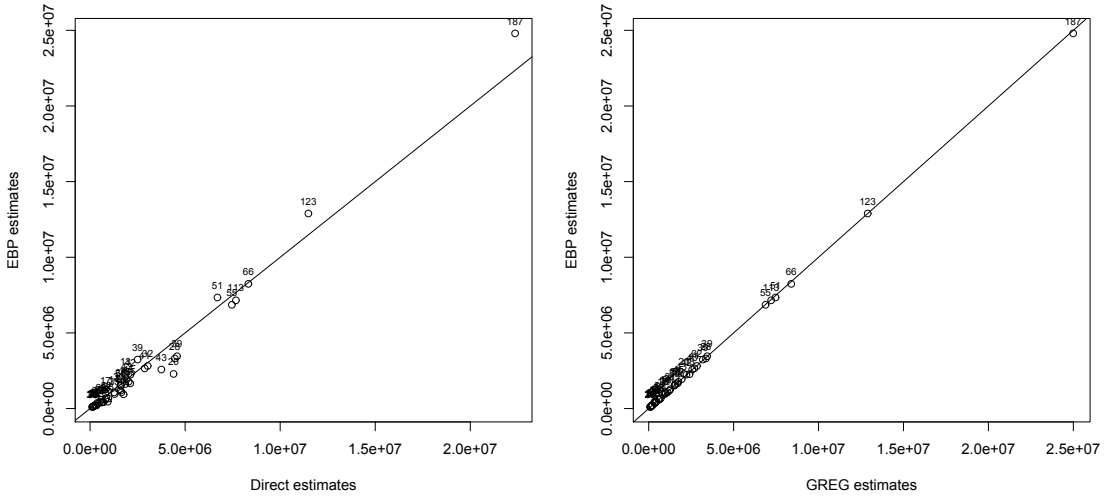


Figure 4.5 (left) plots direct, GREG and EB estimates of total sales for the selected

tobacco product in each Spanish province. This plot shows a great similarity among the considered estimates, with a slightly greater difference for the direct estimates. Figure 4.5 (right) plots the CVs of the three estimators. This figure indicates a better performance of EB estimators in terms of estimated CVs, keeping estimated CV values below 10% for practically all provinces, whereas the GREG shows CV values above 10% for the provinces with the smallest sample sizes. This plot reveals some peaks in the estimated CVs for some provinces with not necessarily the smallest sample sizes. These larger CV values are due to the presence of zero purchases and sales of the considered product in many tobacco shops for those particular provinces. Finally, it is clear that the direct estimator performs the worst in terms of efficiency.

Figure 4.5: Direct, calibration and EB estimates of total sales of tobacco in Spanish provinces (left). Estimated coefficient of variation of direct, calibration and EB estimators of total sales of tobacco in Spanish provinces (right).

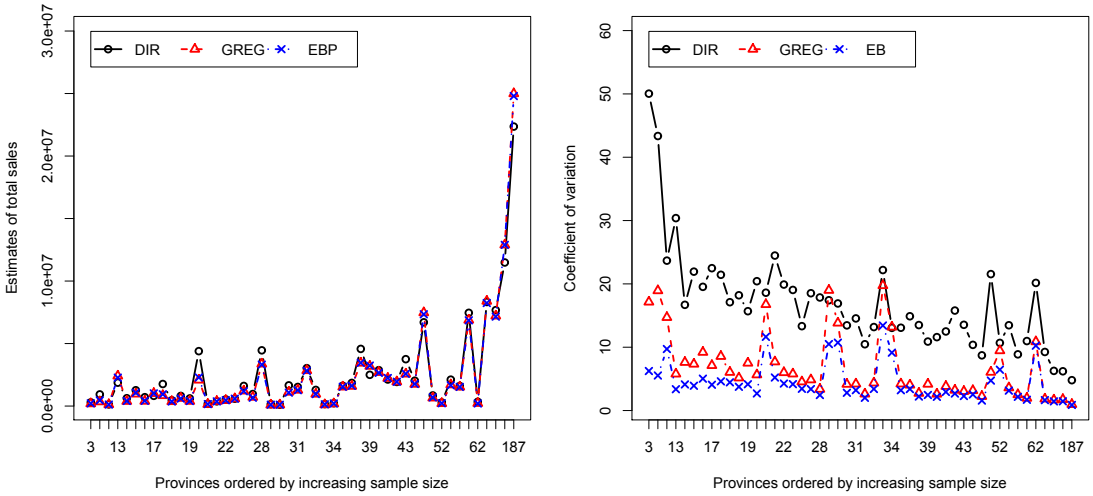


Table C.10 in Appendix C.10 lists the resulting direct, calibration and EB estimates of the particular tobacco product in the Spanish provinces together with the estimated CVs of these estimates. This table confirms that EB achieves the best results in terms of CV specially for those provinces with small sample sizes. As already said, the large CV values for EB are associated with larger percentage of zeros in both sales and purchases of the selected product. Finally, the direct estimator performs poorly in terms of estimated CV even if the bias due to cut-off sampling is estimated not accounted for in his CV.

Chapter 5

Conclusions and future research lines

5.1 Overall conclusions

This Ph. Dissertation focuses on the estimation of general domain parameters in small areas under complex sampling designs, where the selection of the units to the sample depends on their values of the target variable; specifically, under informative sampling and under cut-off sampling.

Due to the important social impact, we illustrate all the procedures through the estimation of poverty indicators typically used by the World Bank. In Chapter 2, we review the most popular poverty mapping procedures, focusing on practical aspects. Simulation studies compare these methods under three interesting scenarios, namely simple random sampling without replacement, informative sampling and presence of outliers in the model. These simulation experiments illustrate the performance of the considered methods when assumptions hold and also when some assumptions are not satisfied. The conclusions that can be drawn are: (i) Aggregation protects against isolated model failures in the FH area level model. Concretely, FH estimates are less affected by symmetric representative unit level outliers and by informative sampling. However, the linearity assumption of the model fails when data follows a unit level model but target parameters are nonlinear functions of the model responses. (ii) EB and HB estimators perform practically the same, and are the best among the considered estimators when the nested error model holds and the sampling design is noninformative. They are not very much affected by mildly informative designs and small proportion of mild outliers, but might be severely affected by highly informative sampling or severe outliers in large proportions. (iii) Census-EB estimators of poverty

indicators are practically the same as EB estimators and avoid linking the survey and census data files. (iv) ELL method under a nested error model with random area effects performs the worst in all scenarios because it does not account for unexplained between-area variation.

In Chapter 3, we focus on informative sampling when estimating additive small area parameters. We propose pseudo EB estimators obtained as expected values with respect to the distribution of out-of-sample variables given the weighted sample means. This method combines the conditioning idea of the EB method for small area estimation of general parameters of Molina and Rao (2010) with the weighting approach of design-based inference. Thus, pseudo EB estimates represent a compromise between model-based and design-based inference. In our simulation studies, pseudo EB estimators reduce considerably the bias of EB estimators when the selection mechanism is informative. On the other hand, under a non-informative sample selection mechanism, the loss of efficiency is small. In the application with Mexican data, we obtained evidences of small upward bias of EB estimates, which could be the result of a partially informative sampling design. This bias seems to be reduced by pseudo EB estimates. The proposed pseudo EB estimates reduce the design bias of purely model-based estimators but, at the same time, gain efficiency with respect to direct estimators. This is achieved with the use of a model representing the common factors that affect the outcomes in all the areas.

Finally, we study an extreme case of informative sampling, which is cut-off sampling. In this design, a set of units of the population are excluded from the possible sample selection and this exclusion is based on their values of the variable of interest. Cut-off sampling is frequently used in business surveys, in which drawing a sample from the whole population entails a high cost that does not compensate the gain in accuracy. On the other hand, in some surveys, part of the target population may not be actually available for sampling; that is, there may be population sectors that are not represented in the sample. These situations appear more often than expected, giving biased direct estimates as we have seen along this work. In Chapter 4, we have studied the design properties of model-based small area estimators under cut-off sampling; precisely, we studied the EBLUP and EB predictor under this setup. We compared the performance of these basic small area estimation methods with the calibration methods proposed by Haziza et al. (2010) for the estimation of small area parameters. Our results indicate that EBLUP or EB perform nearly the same as the calibration estimators in point estimation, reducing the bias due to cut-off sampling. On the other hand, in terms of MSE, EBLUP or EBP perform significantly better than calibration estimators for domains with small sample size.

5.2 Future research lines

In this thesis, we have proposed an effective method to reduce the bias due to informative sampling. Besides, we have studied how small area estimation methods perform under cut-off sampling. Following the research lines in Chapter 3 and Chapter 4, here we comment on the future research lines that we plan to follow.

The method proposed in Chapter 3, pseudo EB, can be extended to more complex linear models using the same general idea of conditioning on weighted sample means. For example, for the two-fold nested error model with domain and subdomain effects, the conditional distribution of the model responses for an out-of-sample unit within a given domain i depends on the (unweighted) sample means of the response variables in each subdomain from that domain. Then, the pseudo EB estimator can be obtained by simply conditioning on the weighted sample means from those subdomains.

The method proposed by Pfeffermann and Sverchkov (2007) for small area estimation under informative selection can also be extended to the estimation of poverty indicators such as the poverty incidence and gap. This method can then be compared with the pseudo EB method proposed here.

Another research line is related with the high computational cost of pseudo EB method, specially when using a bootstrap procedure for estimating the model MSE of pseudo EB estimator. Following the idea of the HB method of Molina et al. (2014), which provides the same point estimates of EB reducing considerably the computational cost, we propose to extend the pseudo EB under the Bayesian approach.

In this dissertation, we have seen that the log-normal distribution does not fit approximately the income, and this is shown by heavier tails in the fitted residuals. We propose to use a much more flexible distribution such as the Generalized Beta of the Second kind (GB2). This distribution fits very well many types of skewed unimodal distributions due to the presence of four parameters instead of the two parameters of the log-normal, with one parameter controlling the length of each tail. It has been effectively used to model income e.g. Graf and Nedyalkova (2014), McDonald and Ransom (2008), Bordley et al. (1996), McDonald (1984), Sepanski and Kong (2008) and McDonald and Xu (1995). We plan to study an estimation approach under a model based on the GB2 distribution that handles data obtained by informative sampling mechanisms or cut-off sampling.

In our simulation studies and the application of Chapter 4, we analyzed the proposed methods assuming that the model is the same for all units in the population (included or excluded). We plan to study the performance of these methods when this is not the case. Note that, when the sampling fraction is negligible and under srswor,

the EBLUP is a convex linear combination of the GREG and the survey regression estimator. Then, if the model is misspecified, we expect that the results will not change significantly, since the EBLUP is expected to become closer to the GREG, which is design-unbiased even if the model fails.

Finally, in Chapter 4, estimated MSEs are obtained for EB or EBLUP under the model, whereas for the direct estimator we have considered the design MSE. Design MSEs are preferred by National Statistical Institutes because they do not assume that a model is correct and therefore account for model failures. There is ongoing research on finding reliable (stable) design MSE estimates of model-based small area estimators, see Strzalkowska and Molina (2017). We plan to use their ideas to find stable design MSE estimators of the considered small area estimators in this context.

Bibliography

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1(7).

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Benedetti, R., Bee, M., and Espa, G. (2010). A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, 26(4):651–671.

Betti, G., Cheli, B., Lemmi, A., and Verma, V. (2006). On the construction of fuzzy measures for the analysis of poverty and social exclusion. *Statistica & Applicazioni*, 4(1):77–97.

Bordley, R., McDonald, J., and Mantrala, A. (1996). Something new, something old: parametric models for the size distribution of income. *Journal of Income Distributions*, 6:91–103.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, 88(423):1013–1020.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro–level estimation of poverty and inequality. *Econometrica*, 71(1):355–364.

Fabrizi, E., Salvati, N., Pratesi, M., and Tzavidis, N. (2014). Outlier robust model-assisted small area estimation. *Biometrical Journal*, 56(1):157–175.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277.

Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica: Journal of the Econometric Society*, pages 761–766.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, 78(5):443–462.

Graf, M. and Nedyalkova, D. (2014). Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. *Review of Income and Wealth*, 60(4):821–842.

Guadarrama, M., Molina, I., and Rao, J. N. K. (2016). A comparison of small area estimation methods for poverty mapping. *Joint issue: Statistics in Transition new series and Survey Methodology*, 17(1):41–66.

Hájek, J. (1971). Discussion of an essay on the logical foundations of survey sampling, part on by d. basu. In Godambe, V. P. and Sprott, D. A., editors, *Foundations of Statistical Inference*, page 326, Toronto. Holt, Rinehart, Winston.

Haziza, D., Chauvet, G., and Deville, J.-C. (2010). Sampling estimation in presence of cut-off sampling. *Australian & New Zealand Journal of Statistics*, 52(3):303–319.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Jiang, J. and Lahiri, P. (2006). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101(473):301–311.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133.

Lehtonen, R., Särndal, C.-E., and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29(1):33–44.

Lehtonen, R. and Veijanen, A. (1999). Domain estimation with logistic generalized regression and related estimators. *Proceedings, IASS Satellite Conference on Small Area Estimation*, pages 121–128.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica: Journal of the Econometric Society*, pages 647–663.

McDonald, J. B. and Ransom, M. (2008). The generalized beta distribution as a model for the distribution of income: estimation of related measures of inequality. *Modeling Income Distributions and Lorenz Curves*, pages 147–166.

McDonald, J. B. and Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics*, 66(1):133–152.

Molina, I. and Marhuenda, Y. (2015). sae: An r package for small area estimation. *R Journal, in print*.

Molina, I. and Morales, D. (2015). Small area estimation of poverty indicators. *Boletin de Estadistica e Investigacion Operativa*, 25:318–325.

Molina, I., Nandram, B., and Rao, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. *The Annals of Applied Statistics*, 8(2):852–885.

Molina, I. and Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3):369–385.

Neri, L., Ballini, F., and Betti, G. (2005). *Poverty and inequality mapping in transition countries*. Università di Siena, Dipartimento di metodi quantitativi.

Pfeffermann, D. et al. (2013). New important developments in small area estimation. *Statistical Science*, 28(1):40–68.

Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102(480):1427–1439.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. *Handbook of Statistics*, 29:455–487.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131.

Prasad, N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85(409):163–171.

Prasad, N. and Rao, J. N. K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, 25:67–72.

Rao, J. N. K. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag.

Sepanski, J. and Kong, J. (2008). A familly of generalized beta distributions for income. *Advances and Application in Statistics*, 10:75–84.

Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, 37(3):381–399.

Stefan, M., Dehon, C., and Droesbeke, J.-J. (2005). Contributions à l'estimation pour petits domaines. *Université libre de Bruxelles*.

Strzalkowska, E. and Molina, I. (2017). Estimation of proportions in small areas: application to the labor force using the swiss census structural survey. *Unpublished work*.

Stubbs, P., Marlier, B., Atkinson, A. B., Cantillon, B., and Nolan, B. (2008). The eu and social inclusion: Facing the challenges. *Journal of Social Policy*, 37:525–526.

Tillé, Y. and Matei, A. (2016). *sampling: Survey Sampling*. R package version 2.8.

Tzavidis, N., Salvati, N., Pratesi, M., and Chambers, R. (2008). M-quantile models with application to poverty mapping. *Statistical Methods & Applications*, 17(3):393–411.

Verret, F., Rao, J. N. K., and Hidiroglou, M. A. (2015). Model-based small area estimation under informative sampling. *Survey Methodology*, 41:333–347.

You, Y. and Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30(3):431–439.

Appendix A

Proofs of Chapter 2

A.1 EBLUP

Under the combined model (2.6) with $F_{\alpha i} = \mathbf{x}_i'\boldsymbol{\beta} + v_i$, the linear estimator $\tilde{F}_{\alpha i} = \alpha_1\hat{F}_{\alpha 1} + \cdots + \alpha_m\hat{F}_{\alpha m}$ that solves the problem:

$$\begin{aligned} \min_{(\alpha_1,\ldots,\alpha_m)} \quad & \text{MSE}_m(\tilde{F}_{\alpha i}) = E(\tilde{F}_{\alpha i} - F_{\alpha i})^2 \\ \text{s.t.} \quad & E_m(\tilde{F}_{\alpha i} - F_{\alpha i}) = 0 \end{aligned}$$

is given by $\tilde{F}_{\alpha i}^{BLUP} = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{v}_i$, where $\tilde{v}_i = \tilde{v}_i(\sigma_v^2) = \gamma_i(\hat{F}_{\alpha i} - \mathbf{x}_i'\tilde{\boldsymbol{\beta}})$, $\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$ and

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_v^2) = \left(\sum_{i=1}^m \gamma_i \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^m \gamma_i \mathbf{x}_i \hat{F}_{\alpha i}.$$

Proof: We prove a more general result. Let us express model (2.6) in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} + \mathbf{e},$$

where

$$\mathbf{y} = \begin{pmatrix} \hat{F}_{\alpha 1} \\ \vdots \\ \hat{F}_{\alpha m} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_m' \end{pmatrix}, \mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_m \end{pmatrix}.$$

The covariance matrices of the vectors of area effects and errors are: $V_m(\mathbf{v}) = \sigma_v^2\mathbf{I}_m$ and $V_m(\mathbf{e}) = \text{diag}_{1\le i\le m}(\psi_i)$ respectively.

We prove that the best linear unbiased predictor (BLUP) of the mixed effect $\mu = \boldsymbol{\ell}'\boldsymbol{\beta} + \mathbf{t}'\mathbf{v}$, for given $p \times 1$ and $m \times 1$ vectors of known constants $\boldsymbol{\ell}$ and $\mathbf{t}$, is given by $\tilde{\mu} = \boldsymbol{\ell}'\tilde{\boldsymbol{\beta}} + \mathbf{t}'\tilde{\mathbf{v}}$, for $\tilde{\mathbf{v}} = (\tilde{v}_1, \ldots, \tilde{v}_m)'$. A linear predictor of $\mu = \boldsymbol{\ell}'\boldsymbol{\beta} + \mathbf{t}'\mathbf{v}$ is given

by $\tilde{\mu} = \boldsymbol{\alpha}'\mathbf{y} + b$, for a given vector of constants $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ and scalar b. The prediction error is then

$$\tilde{\mu} - \mu = \boldsymbol{\alpha}'\mathbf{y} + b - \boldsymbol{\ell}'\boldsymbol{\beta} - \mathbf{t}'\mathbf{v} = \boldsymbol{\alpha}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\alpha}'\mathbf{v} + \boldsymbol{\alpha}'\mathbf{e} + b - \boldsymbol{\ell}'\boldsymbol{\beta} - \mathbf{t}'\mathbf{v}.$$

We say that $\tilde{\mu}$ is model unbiased for μ iff $E_m(\tilde{\mu}-\mu) = 0$. Then, taking expected value of the above prediction error, we obtain $E_m(\tilde{\mu}-\mu) = (\boldsymbol{\alpha}'\mathbf{X} - \boldsymbol{\ell}')\boldsymbol{\beta} + b$, which equals zero $\forall\boldsymbol{\beta}$ iff $\boldsymbol{\alpha}'\mathbf{X} = \boldsymbol{\ell}'$ and $b = 0$. Now, if $\tilde{\mu}$ is unbiased for μ, then the MSE of $\tilde{\mu}$ equals the variance of the prediction error, that is,

$$\text{MSE}_m(\tilde{\mu}) = V_m(\tilde{\mu} - \mu) = V_m(\boldsymbol{\alpha}'\mathbf{y} - \mathbf{t}'\mathbf{v}) = \boldsymbol{\alpha}'\mathbf{V}\boldsymbol{\alpha} + \sigma_v^2\mathbf{t}'\mathbf{t} - 2\sigma_v^2\boldsymbol{\alpha}'\mathbf{t},$$

where $\mathbf{V} = V_m(\mathbf{y}) = \sigma_v^2\mathbf{I}_m + \text{diag}(\psi_i)$. Thus, the BLUP solves the following minimization problem

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \text{MSE}_m(\tilde{\mu}) = \boldsymbol{\alpha}'\mathbf{V}\boldsymbol{\alpha} + \sigma_v^2\mathbf{t}'\mathbf{t} - 2\sigma_v^2\boldsymbol{\alpha}'\mathbf{t} \\ \text{s.t.} \quad & \boldsymbol{\alpha}'\mathbf{X} = \boldsymbol{\ell}'. \end{aligned}$$

By the Lagrange multiplier method, we obtain

$$\boldsymbol{\alpha}' = \boldsymbol{\ell}'(\mathbf{X}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \sigma_v^2\mathbf{t}'\mathbf{V}^{-1}[\mathbf{I}_m - \mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}].$$

Replacing in $\tilde{\mu} = \boldsymbol{\alpha}'\mathbf{y}$ and calling $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, the BLUP of μ is given by

$$\tilde{\mu}^{BLUP} = \boldsymbol{\alpha}'\mathbf{y} = \boldsymbol{\ell}'\tilde{\boldsymbol{\beta}} + \sigma_v^2\mathbf{t}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \boldsymbol{\ell}'\tilde{\boldsymbol{\beta}} + \mathbf{t}'\tilde{\mathbf{v}},$$

where $\tilde{\mathbf{v}} = \sigma_v^2\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$. Finally, taking $\boldsymbol{\ell} = \mathbf{x}_i$ and $\mathbf{t}$ as an $m \times 1$ vector of zeros with 1 in position i, that is, $\mathbf{t} = (0, \ldots, 0, 1, 0, \ldots, 0)$, we obtain

$$\tilde{F}_i^{BLUP} = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{v}_i.$$

A.2 Posterior distribution for HB

In this section, we obtain the posterior distribution of $\boldsymbol{\theta} = (\mathbf{v}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$. This distribution is proper under certain regularity conditions (see Molina et al. (2014)) and is given by

$$\pi(\mathbf{v}, \boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s) = \pi_1(\mathbf{v}|\boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s)\pi_2(\boldsymbol{\beta}|\sigma_e^2, \rho, \mathbf{y}_s), \pi_3(\sigma_e^2|\rho, \mathbf{y}_s)\pi_4(\rho|\mathbf{y}_s).$$

Now to obtain the conditional distributions, $\pi_1 - \pi_4$, we proceed as follows: π_1 is the result of integrating out $\mathbf{v}$ from $\pi(\mathbf{v}, \boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s)$ and dividing $\pi(\mathbf{v}, \boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s)$ by $\pi(\boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s)$, that is, we first calculate

$$\pi(\boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s) \;=\; \int_{-\infty}^{\infty} \pi(\mathbf{v}, \boldsymbol{\beta}, \sigma_e^2 \rho|\mathbf{y}_s) d\mathbf{v},$$

where

$$\pi(\mathbf{v}, \boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s) = \left(\frac{1-\rho}{\rho}\right)^{m/2} (\sigma_e^2)^{-\left(\frac{m+n}{2}+1\right)} \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{m} \left[\sum_{j \in s_i} w_{ij}(Y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - v_i)^2 + \frac{1-\rho}{\rho} v_i^2 \right] \right\},$$

and, then, we obtain

$$\pi_1(\mathbf{v}|\boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s) \;=\; \frac{\pi(\mathbf{v}, \boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s)}{\pi(\boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s)} = \prod_{i=1}^{m} \pi(v_i|\boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s),$$

where

$$v_i|\boldsymbol{\beta}, \sigma_e^2, \rho, \mathbf{y}_s \overset{ind}{\sim} N\left[\lambda_d(\rho)(\bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta}), \{1 - \lambda_i(\rho)\} \frac{\rho}{1-\rho} \sigma_e^2 \right],$$

for $\lambda_i(\rho) = w_{i\cdot}[w_{i\cdot} + (1-\rho)/\rho]^{-1}$, $i = 1, \ldots, m$ and $w_{i\cdot} = \sum_{j \in s_i} w_{ij}$. To obtain π_2, we integrate out $\boldsymbol{\beta}$ from $\pi(\boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s)$ and, then, we divide $\pi(\boldsymbol{\beta}, \sigma_e^2, \rho|\mathbf{y}_s)$ by $\pi(\sigma_e^2, \rho|\mathbf{y}_s)$. Then $\pi_2(\boldsymbol{\beta}|\sigma_e^2, \rho, \mathbf{y}_s)$ turns out to be

$$\boldsymbol{\beta}|\sigma_e^2, \rho, \mathbf{y}_s \sim N\left\{ \hat{\boldsymbol{\beta}}(\rho), \sigma_e^2 \mathbf{Q}^{-1}(\rho) \right\},$$

where $\hat{\boldsymbol{\beta}}(\rho) = \mathbf{Q}^{-1}(\rho)\mathbf{p}(\rho)$, and $\mathbf{Q}^{-1}(\rho)$ and $\mathbf{p}(\rho)$ are given respectively by

$$\begin{aligned} \mathbf{Q}(\rho) &= \sum_{i=1}^{m} \sum_{j \in s_i} w_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' + \frac{1-\rho}{\rho} \sum_{i=1}^{m} \lambda_i(\rho)\, \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i', \\ \mathbf{p}(\rho) &= \sum_{di=1}^{m} \sum_{j \in s_i} w_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(Y_{ij} - \bar{y}_i) + \frac{1-\rho}{\rho} \sum_{i=1}^{m} \lambda_i(\rho) \bar{\mathbf{x}}_i \bar{y}_i, \end{aligned}$$

for $\bar{\mathbf{x}}_i = w_{i\cdot} \sum_{j \in s_i} w_{ij} x_{ij}$ and $\bar{y}_i = w_{i\cdot} \sum_{j \in s_i} w_{ij} y_{ij}$. The posterior densities π_4 and π_3 are obtained respectively as follows,

$$\pi_4(\rho|\mathbf{y}_s) \quad = \quad \int_{-\infty}^{\infty} \pi(\sigma_e^2, \rho|\mathbf{y}_s) d\sigma_e^2,$$

which is proportional to

$$\pi_4(\rho|\mathbf{y}_s) \propto \left(\frac{1-\rho}{\rho}\right)^{m/2} |\mathbf{Q}(\rho)|^{-1/2} \gamma(\rho)^{-(n-p)/2} \prod_{i=1}^{m} \lambda_i^{1/2}(\rho), \ \epsilon \leq \rho \leq 1-\epsilon,$$

for $\gamma(\rho)$ given by

$$\gamma(\rho) = \sum_{i=1}^{m} \sum_{j \in s_i} w_{ij} \left\{ Y_{ij} - \bar{y}_i - (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' \hat{\boldsymbol{\beta}}(\rho) \right\}^2 + \frac{1-\rho}{\rho} \sum_{i=1}^{m} \lambda_i(\rho) \left\{ \bar{y}_i - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}(\rho) \right\}^2.$$

Finally, π_3 is given by

$$\pi_3(\sigma_e^2|\rho, \mathbf{y}_s) \quad = \quad \frac{\pi(\sigma_e^2, \rho|\mathbf{y}_s)}{\pi_4(\rho|\mathbf{y}_s)},$$

which results

$$\sigma_e^{-2}|\rho, \mathbf{y}_s \sim \text{Gamma}\left(\frac{n-p}{2}, \frac{\gamma(\rho)}{2}\right).$$

Appendix B

Proofs of Chapter 3

In this appendix, we prove the design consistency and asymptotic unbiasedness of the Census Pseudo Best predictors of domain poverty incidence and gap. For simplicity of exposition, we give the formal results for the case of area-level covariates, that is, $\mathbf{x}_{ij} = \mathbf{x}_i$, where $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. Let us define $\bar{\epsilon}_{iw} = \bar{y}_{iw} - \mathbf{x}_i'\boldsymbol{\beta}$. Consider that the model (3.1)-(3.2) holds for $Y_{ij} = \log(E_{ij} + c)$ as in Section 3.3. According to (3.17) with $H_i = F_{\alpha i}$ and using (3.10) with conditional means and variances given in (3.16), the Census Pseudo Best predictor of the poverty incidence F_{0i} can be expressed as

$$\tilde{F}_{0i}^{CPB} = \Phi\left(\frac{\log(z+c) - \mathbf{x}_i'\boldsymbol{\beta} - \gamma_{iw}\bar{\epsilon}_{iw}}{\sqrt{\sigma_v^2(1-\gamma_{iw}) + \sigma_e^2}}\right) = g_0(\gamma_{iw}, \bar{\epsilon}_{iw}). \tag{B.1}$$

Similarly, we can express the Census Pseudo Best predictor of the poverty gap, F_{1i}, as

$$\begin{aligned}\tilde{F}_{1i}^{CPB} &= \Phi\left(\frac{\log(z+c) - \mathbf{x}_i'\boldsymbol{\beta} - \gamma_{iw}\bar{\epsilon}_{iw}}{\sqrt{\sigma_v^2(1-\gamma_{iw}) + \sigma_e^2}}\right)\left\{1 - \frac{1}{z}\left[e^{\mathbf{x}_i'\boldsymbol{\beta} - \gamma_{iw}\bar{\epsilon}_{iw} + \frac{\sigma_v^2(1-\gamma_{iw}) + \sigma_e^2}{2}}\right.\right. \\ &\left.\left. \times \frac{\Phi\left(\frac{\log(z+c) - \mathbf{x}_i'\boldsymbol{\beta} - \gamma_{iw}\bar{\epsilon}_{iw}}{\sqrt{\sigma_v^2(1-\gamma_{iw}) + \sigma_e^2}} - \sqrt{\sigma_v^2(1-\gamma_{iw}) + \sigma_e^2}\right)}{\Phi\left(\frac{\log(z+c) - \mathbf{x}_i'\boldsymbol{\beta} - \gamma_{iw}\bar{\epsilon}_{iw}}{\sqrt{\sigma_v^2(1-\gamma_{iw}) + \sigma_e^2}}\right)} - c\right]\right\} \\ &= g_1(\gamma_{iw}, \bar{\epsilon}_{iw}).\end{aligned} \tag{B.2}$$

Let us define $\bar{\xi}_i = \bar{Y}_i - \mathbf{x}_i'\boldsymbol{\beta}$, $F_{\alpha i}^* = g_\alpha(1, \bar{\xi}_i)$, $\alpha = 0, 1$, $W_i = \sum_{j=1}^{N_i} w_{ij}$ and $\bar{W}_i = N_i^{-1} W_i$. Proposition 1 gives the design consistency of $\tilde{F}_{\alpha i}^{CPB}$ for $F_{\alpha i}^*$ under general sampling designs.

Proposition 1. *Under the regularity assumptions required for design consistency of the Horvitz-Thompson estimators $\hat{Y}_i = \sum_{j \in s_i} w_{ij} Y_{ij}$, $\hat{W}_i = \sum_{j \in s_i} w_{ij}^2$ and $\hat{N}_i = \sum_{j \in s_i} w_{ij}$ of the totals $Y_i = \sum_{j=1}^{N_i} Y_{ij}$, $W_i = \sum_{j=1}^{N_i} w_{ij}$ and N_i (see e.g. Isaki and Fuller (1982)), if additionally*

$N_i^{-1}\bar{W}_i \to 0$ *as* $n_i \to \infty$ *and* $N_i \to \infty$, *then*

$$|\tilde{F}_{\alpha i}^{CPB} - F_{\alpha i}^*| \xrightarrow[n_i\to\infty]{P_\pi} 0,\ \alpha = 0, 1,$$

where P_π is the probability distribution under the sampling replication mechanism.

Remark B.1. In fact, the same regularity assumptions of Isaki and Fuller (1982) give consistency in quadratic mean with respect to the design, P_π, that is

$$|\tilde{F}_{\alpha i}^{CPB} - F_{\alpha i}^*| \xrightarrow[n_i\to\infty]{} 0 \text{ in quadratic mean.} \tag{B.3}$$

Proof: By the considered assumptions, as $n_i \to \infty$ and $N_i \to \infty$, it holds $\bar{y}_{iw} = \hat{Y}_i/\hat{N}_i \xrightarrow[n_i\to\infty]{P_\pi} \bar{Y}_i$, which implies $\bar{\epsilon}_{iw} \xrightarrow[n_i\to\infty]{P_\pi} \bar{\xi}_i$. Moreover, we have $\delta_i^2 = \hat{W}_i/\hat{N}_i^2 \xrightarrow[n_i\to\infty]{P_\pi} \bar{W}_i/N_i \to 0$ as $n_i \to \infty$, which implies that $\gamma_{iw} \xrightarrow[n_i\to\infty]{P_\pi} 1$. The result then follows by the continuity of the functions $g_\alpha(\cdot,\cdot)$, $\alpha = 0, 1$, defining the CPB predictors (B.1) and (B.2). The conditions of Isaki and Fuller (1982) provide in fact consistency in quadratic mean (B.3). □

In what follows, we consider the model for $Y_{ij}|v_i$ given in (3.1) and we call it model m_1, and we make no distributional assumptions for v_i unlike the full model (3.1)-(3.2). We use E_{m_1} to denote expectation under this model. Proposition 2 gives the expectation of the poverty incidence and gap under model m_1.

Proposition 2. *The expectation of* $F_{\alpha i}$ *under the model for* $Y_{ij}|v_i$ *given in* (3.1) *(called model* m_1*) is given by*

$$E_{m_1}(F_{\alpha i}) = g_\alpha(1, v_i),\ \alpha = 0, 1. \tag{B.4}$$

The next result gives asymptotic unbiasedness of the CPB predictors of $F_{\alpha i}$, for $\alpha = 0, 1$, with respect to the design and the model m_1.

Proposition 3. *Under the regularity assumptions required for design consistency of the Horvitz-Thompson estimators* $\hat{Y}_i = \sum_{j\in s_i} w_{ij}Y_{ij}$, $\hat{W}_i = \sum_{j\in s_i} w_{ij}^2$ *and* $\hat{N}_i = \sum_{j\in s_i} w_{ij}$ *of the totals* $Y_i = \sum_{j=1}^{N_i} Y_{ij}$, $W_i = \sum_{j=1}^{N_i} w_{ij}$ *and* N_i *(see e.g. Isaki and Fuller, 1982), if additionally* $N_i^{-1}\bar{W}_i \to 0$ *as* $n_i \to \infty$, *then*

$$E_{m_1}E_\pi(\tilde{F}_{\alpha i}^{CPB}) - E_{m_1}(F_{\alpha i}) \to 0,\ \alpha = 0, 1,\ \text{as } n_i \to \infty \text{ and } N_i \to \infty.$$

Proof: Under the mentioned regularity conditions, by (B.3), we have

$$E_\pi(\tilde{F}_{\alpha i}^{CPB}) - F_{\alpha i}^* \to 0 \text{ as } n_i \to \infty. \tag{B.5}$$

Moreover, under model m_1, we have $\bar{E}_i = N_i^{-1} \sum_{j=1}^{N_i} e_{ij} \xrightarrow[N_i \to \infty]{a.s} 0 = E_{m_1}[e_{ij}]$. Then,

$$\bar{\xi}_i - v_i = \bar{E}_i \xrightarrow[N_i \to \infty]{a.s.} 0,$$

which, by continuity of functions $g_\alpha(\cdot, \cdot)$ for $\alpha = 0, 1$, implies that

$$g_\alpha(1, \bar{\xi}_i) - g_\alpha(1, v_i) \xrightarrow[N_i \to \infty]{a.s.} 0.$$

By (B.4), we get

$$F^*_{\alpha i} - E_{m_1}(F_{\alpha i}) \xrightarrow[N_i \to \infty]{a.s.} 0,$$

and noting that a.s. convergence implies convergence in mean, we obtain

$$E_{m_1}(F^*_{\alpha i}) - E_{m_1}(F_{\alpha i}) \to 0. \tag{B.6}$$

Finally, by (B.5) and (B.6), we get

$$E_{m_1} E_\pi(\tilde{F}^{CPB}_{\alpha i}) - E_{m_1}(F_{\alpha i}) = E_{m_1}\{E_\pi(\tilde{F}^{CPB}_{\alpha i}) - F^*_{\alpha i}\} + E_{m_1}(F^*_{\alpha i}) - E_{m_1}(F_{\alpha i}) \to 0,$$

as $n_i \to \infty$ and $N_i \to \infty$. □

Remark B.2. Under simple random sampling within domain i, we have $N_i^{-1} \bar{W}_i = n_i^{-1} \to 0$, which implies that the Census Best predictors $\tilde{F}^{CB}_{\alpha i}$ obtained with the true $\boldsymbol{\theta}$ are asymptotically unbiased when taking expectation under that sampling mechanism and under m_1 as $n_i \to \infty$ and $N_i \to \infty$.

The above results can be extended to the case of non-constant unit level vectors $\mathbf{x}_{ij}$ by writing $\tilde{F}^{CPB}_{\alpha i} = N_i^{-1} \sum_{j=1}^{N_i} g_{\alpha j}(\gamma_{iw}, \bar{\epsilon}_{iw})$ and $F^*_{\alpha i} = N_i^{-1} \sum_{j=1}^{N_i} g_{\alpha j}(1, \bar{\xi}_i)$, and making a Taylor expansion of $\tilde{F}^{CPB}_{\alpha i} = N_i^{-1} \sum_{j=1}^{N_i} g_{\alpha j}(\gamma_{iw}, \bar{\epsilon}_{iw})$ around $(1, \bar{\xi}_i)$. Further regularity assumptions are required.

Appendix C

Proofs of Chapter 4

C.1 Equality of basic calibration estimator and GREG

We prove that $\hat{Y}^{LCAL} = \hat{Y}^{GREG}$ when $G_j(h, w)$ is the chi-squared distance. By the definition of the LCAL estimator in (4.6), and using the expression of the calibrated weights h_j and $\boldsymbol{\lambda}$ given in (4.4) and (4.5), we obtain

$$\begin{aligned}
\hat{Y}^{LCAL} &= \sum_{j\in s} h_j y_j \\
&= \sum_{j\in s} w_j(1 + \mathbf{x}_j'\boldsymbol{\lambda})y_j \\
&= \sum_{j\in s}(w_j y_j + w_j \mathbf{x}_j'\boldsymbol{\lambda} y_j) \\
&= \sum_{j\in s}(w_j y_j + w_j \boldsymbol{\lambda}'\mathbf{x}_j y_j) \\
&= \sum_{j\in s} w_j y_j + (\mathbf{X} - \hat{\mathbf{X}})' \left(\sum_{j\in s} w_j \mathbf{x}_j \mathbf{x}_j'\right)^{-1} \sum_{j\in s} w_j \mathbf{x}_j y_j \\
&= \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})'\hat{\mathbf{B}} \\
&= \hat{Y}^{GREG}.
\end{aligned}$$

C.2 Derivation of calibration function $F(\cdot)$

Define $G_j(h, w)$ such that,

(1) $\forall w > 0$, $G_j(h, w) \geqslant 0$, $\partial G_j(h, w)/\partial h$ exists, $G_j(h, w) \geqslant 0$ is strictly convex in h, for h in an interval $D_j(w)$ containing w with $G_j(w, w) = 0$.

(2) The function $g_j(h,w) = \partial G_j(h,w)/\partial h$ is continuous and $g_j(\cdot,w)$ maps $D_j(w)$ onto an interval $\text{Im}_j(w)$ in a one-to-one fashion.

By (1), $G_j(h,w) = 0 \Leftrightarrow h = w$ is the minimum of G_j. Then,

$$\partial G_j(h,w)/\partial h = g_j(h,w) = 0 \text{ iff } h = w \text{ and } g_j(w,w) = 0.$$

Calibrated weights h_j are obtained by solving the problem

$$\begin{aligned} \min \quad & \sum_{j\in s} G_j(h,w) \\ \text{s.t.} \quad & \sum_{j\in s} h_j \mathbf{x}_j = \mathbf{X}. \end{aligned}$$

By Lagrange multipliers' method

$$L = \sum_{j\in s} G_j(h,w) + 2\boldsymbol{\lambda}' \left(\sum_{j\in s} h_j \mathbf{x}_j - \mathbf{X} \right),$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. Taking the derivative of L with respect to the new weights h_j and equating to zero, we obtain

$$\begin{aligned} \frac{\partial L}{\partial h_j} &= \frac{\partial G_j(h,w)}{\partial h_j} - \boldsymbol{\lambda}'\mathbf{x}_j = 0 \\ &\Leftrightarrow \quad g_j(h,w) - \boldsymbol{\lambda}'\mathbf{x}_j = 0 \\ &\Leftrightarrow \quad g_j(h,w) = \boldsymbol{\lambda}'\mathbf{x}_j \\ &\Leftrightarrow \quad h_j = g_j^{-1}(\mathbf{x}_j'\boldsymbol{\lambda}), \end{aligned}$$

where $g_j^{-1}(\cdot)$ is the inverse fucntion of $g_j(\cdot,w)$. Thus, we have $F(\cdot) = g_j^{-1}(\cdot)/w_j$ and this function satisfies $F(0) = g_j^{-1}(0)/w_j = w_j/w_j = 1$, because $g_j^{-1}(0) = w_j$, since $g_j(h,w) = 0 \Leftrightarrow h_j = w_j$.

C.3 Derivation of LCALN estimator

The Lagrangian function corresponding to the minimization problem (4.17) is

$$L = \sum_{i=1}^{m} \sum_{j\in s_i} (h_{ij} - w_{ij})^2/w_{ij} + 2\boldsymbol{\lambda}' \left(\sum_{i=1}^{m} \sum_{j\in s_i} h_{ij}\mathbf{x}_{ij} - \mathbf{X} \right).$$

Taking derivatives of L with respect to h_{ij} and equating to zero, we obtain

$$\begin{aligned}\frac{\partial L}{\partial h_{ij}} &= 2w_{ij}(h_{ij}/w_{ij}-1)/w_{ij} - 2\boldsymbol{\lambda}'\mathbf{x}_{ij} = 0\\ &\Leftrightarrow\ h_{ij}/w_{ij} - 1 = \boldsymbol{\lambda}'\mathbf{x}_{ij}\\ &\Leftrightarrow\ h_{ij} = w_{ij}(1+\mathbf{x}_{ij}'\boldsymbol{\lambda}).\end{aligned}$$

On the other hand, taking derivatives of L with respect to $\boldsymbol{\lambda}$, we obtain

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = 2\left(\sum_{i=1}^{m}\sum_{j\in s_i} h_{ij}\mathbf{x}_{ij} - \mathbf{X}\right) = \mathbf{0}_p.$$

Replacing $h_{ij} = w_{ij}(1+\mathbf{x}_{ij}'\boldsymbol{\lambda})$ from the previous equation in the latter, we obtain

$$\begin{aligned}&\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}(1+\mathbf{x}_{ij}'\boldsymbol{\lambda})\mathbf{x}_{ij} - \mathbf{X} = \mathbf{0}_p\\ \Leftrightarrow\ &\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij} + \sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}'\boldsymbol{\lambda} - \mathbf{X} = \mathbf{0}_p\\ \Leftrightarrow\ &\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}'\boldsymbol{\lambda} = \mathbf{X} - \sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\\ \Leftrightarrow\ &\boldsymbol{\lambda} = \left(\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}'\right)^{-1}\left(\mathbf{X} - \sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\right)\\ \Leftrightarrow\ &\boldsymbol{\lambda} = \left(\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}'\right)^{-1}\left(\mathbf{X} - \hat{\mathbf{X}}\right),\end{aligned}$$

provided that $\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}'$ is non-singular, where $\hat{\mathbf{X}} = \sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}$.

C.4 Design-bias of LCAL estimator under cut-off sampling

We calculate the bias under the design of the theoretical LCAL estimator for a domain total given in (4.22). Taking expectation of (4.22), we obtain

$$\begin{aligned}E_\pi(\tilde{Y}_i) &= E_\pi[\hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)'\mathbf{B}_{iI}]\\ &= Y_{iI} + (\mathbf{X}_i - \mathbf{X}_{iI})'\mathbf{B}_{iI}.\end{aligned}$$

Therefore, the design bias of $\tilde{Y}_i^{LCAL}$ is given by

$$\begin{aligned} B_\pi(\tilde{Y}_i^{LCAL}) &= E_\pi(\tilde{Y}_i^{LCAL}) - Y_i \\ &= Y_{iI} + (\mathbf{X}_i - \mathbf{X}_{iI})'\mathbf{B}_{iI} - Y_i \\ &= Y_{iI} + \mathbf{X}'_{iE}\mathbf{B}_{iI} - Y_{iI} - Y_{iE} \\ &= -(Y_{iE} - \mathbf{X}'_{iE}\mathbf{B}_{iI}). \end{aligned}$$

For the theoretical LCAL estimator of the domain mean, $\tilde{\bar{Y}}_i$ given in (4.25), the expectation under the design is given by

$$\begin{aligned} E_\pi(\tilde{\bar{Y}}_i^{LCAL}) &= E_\pi(\hat{\bar{Y}}_i) + E_\pi[(\bar{\mathbf{X}}_i - \hat{\bar{\mathbf{X}}}_i)'\mathbf{B}_{iI}] \\ &= \bar{Y}_{iI} + (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_{iI})'\mathbf{B}_{iI}. \end{aligned}$$

Therefore, the bias under the design of $\tilde{\bar{Y}}_i^{LCAL}$ is given by

$$\begin{aligned} B_\pi(\tilde{\bar{Y}}_i^{LCAL}) &= E_\pi(\tilde{\bar{Y}}_i^{LCAL}) - \bar{Y}_i \\ &= \bar{Y}_{iI} + (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_{iI})'\mathbf{B}_{iI} - \bar{Y}_i \\ &= \bar{Y}_{iI} + (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_{iI})'\mathbf{B}_{iI} - \frac{N_{iI}\bar{Y}_{iI} + N_{iE}\bar{Y}_{iE}}{N_i} \\ &= \frac{-N_{iI}\bar{Y}_{iI}}{N_i} + \bar{Y}_{iI} + (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_{iI})'\mathbf{B}_{iI} - \frac{N_{iE}\bar{Y}_{iE}}{N_i} \\ &= \frac{N_{iE}}{N_i}\bar{Y}_{iI} + (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_{iI})'\mathbf{B}_{iI} - \frac{N_{iE}}{N_i}\bar{Y}_{iE} \\ &= \frac{N_{iE}}{N_i}[(\bar{Y}_{iI} - \bar{\mathbf{X}}'_{iI}\mathbf{B}_{iI}) - (\bar{Y}_{iE} - \bar{\mathbf{X}}'_{iE}\mathbf{B}_{iI})]. \end{aligned}$$

The variance under the design of the theoretical LCAL estimator (4.25) is given by

$$\begin{aligned} V_\pi(\tilde{Y}_i^{LCAL}) &= V_\pi[\hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)'\mathbf{B}_{iI}] = V_\pi(\hat{Y}_i - \hat{\mathbf{X}}'_i\mathbf{B}_{iI}) \\ &= V_\pi\left(\sum_{j\in s_i} w_{ij}y_{ij} - \sum_{j\in s_i} w_{ij}\mathbf{x}'_{ij}\mathbf{B}_{iI}\right) \\ &= V_\pi\left[\sum_{j\in s_i} w_{ij}(y_{ij} - \mathbf{x}'_{ij}\mathbf{B}_{iI})\right] \\ &= V_\pi\left(\sum_{j\in s_i} w_{ij}\varepsilon_{ij}\right) \end{aligned}$$

for $\varepsilon_{ij} = y_{ij} - \mathbf{x}'_{ij}\mathbf{B}_{iI}$. using the basic design-based inference, let $I_{ij}(s)$ be the indicator of unit j from U_{iI} belonging to the sample s_i, that is

$$I_{ij}(s) = \begin{cases} 1 & \text{if } j \in s_i; \\ 0 & \text{if } j \notin s_i; \quad j \in U_{iI}. \end{cases}$$

The design variance of the HT estimator $\sum_{j\in s_i} w_{ij}\varepsilon_{ij}$ of the total $\sum_{j\in U_{iI}} \varepsilon_{ij}$ is then

$$\begin{aligned} V_\pi\left(\sum_{j\in s_i} w_{ij}\varepsilon_{ij}\right) &= \sum_{j\in U_{iI}} w_{ij}^2\varepsilon_{ij}^2 V_\pi(I_{ij}(s)) + 2\sum_{j\in U_{iI}}\sum_{k\in U_{iI}} \varepsilon_{ij}\varepsilon_{ik} Cov_\pi[I_{ij}(s), I_{ik}(s)] \\ &= \sum_{j\in U_{iI}} w_{ij}^2\varepsilon_{ij}^2\pi_{ij}(1-\pi_{ij}) + 2\sum_{j\in U_{iI}}\sum_{k\in U_{iI}} \frac{\varepsilon_{ij}\varepsilon_{ik}}{\pi_{ij}\pi_{ik}}\left(\pi_{jk}^i - \pi_{ij}\pi_{ik}\right) \\ &= \sum_{j\in U_{iI}} w_{ij}^2\varepsilon_{ij}^2\frac{1}{w_{ij}}\left(1-\frac{1}{w_{ij}}\right) + 2\sum_{j\in U_{iI}}\sum_{k\in U_{iI}} \varepsilon_{ij}\varepsilon_{ik} w_{ij}w_{ik}\left(\frac{w_{ij}w_{ik} - w_{jk}^i}{w_{jk}^i w_{ij}w_{ik}}\right) \\ &= \sum_{j\in U_{iI}} \varepsilon_{ij}^2(w_{ij}-1) + 2\sum_{j\in U_{iI}}\sum_{k\in U_{iI}} \varepsilon_{ij}\varepsilon_{ik}\left(\frac{w_{ij}w_{ik} - w_{jk}^i}{w_{jk}^i}\right). \end{aligned}$$

Note that the second term of the variance is approximately equal to zero whenever $\pi_{ij}\pi_{ik} \cong \pi_{jk}^i$. Finally, for the area mean $\bar{Y}_i$, the variance under the design of $\tilde{\bar{Y}}_i^{LCAL}$ is given by

$$V_\pi(\tilde{\bar{Y}}_i^{LCAL}) = \frac{1}{N_i^2}\left[\sum_{j\in U_{iI}} \varepsilon_{ij}^2(w_{ij}-1) + 2\sum_{j\in U_{iI}}\sum_{k\in U_{iI}} \varepsilon_{ij}\varepsilon_{ik}\left(\frac{w_{ij}w_{ik} - w_{jk}^i}{w_{jk}^i}\right)\right].$$

C.5 Design-bias of LCALN estimator under cut-off sampling

We derive the bias under the design of the theoretical LCALN estimator of the domain mean. The design-bias of (4.30) is then given by

$$\begin{aligned} B_\pi(\tilde{\bar{Y}}_i^{LCALN}) &= E_\pi(\tilde{\bar{Y}}_i^{LCALN}) - \bar{Y}_i \\ &= \bar{Y}_{iI} + \left(\frac{\mathbf{X}}{N} - \frac{\mathbf{X}_I}{N_I}\right)' \mathbf{B}_{iI}^N - \frac{N_{iI}\bar{Y}_{iI} + N_{iE}\bar{Y}_{iE}}{N_i} \\ &= \frac{N_{iE}}{N_i}\bar{Y}_{iI} + \left(\frac{N_I\mathbf{X}_I + N_I\mathbf{X}_E - N\mathbf{X}_I}{NN_I}\right)' \mathbf{B}_{iI}^N - \frac{N_{iE}}{N_i}\bar{Y}_{iE} \\ &= \frac{N_{iE}}{N_i}\bar{Y}_{iI} + \left(\frac{N_E}{N}\bar{\mathbf{X}}_E - \frac{N_E}{N}\bar{\mathbf{X}}_I\right)' \mathbf{B}_{iI}^N - \frac{N_{iE}}{N_i}\bar{Y}_{iE} \\ &= \left(\frac{N_{iE}}{N_i}\bar{Y}_{iI} - \frac{N_E}{N}\bar{\mathbf{X}}_I'\mathbf{B}_{iI}^N\right) - \left(\frac{N_{iE}}{N_i}\bar{Y}_{iE} - \frac{N_E}{N}\bar{\mathbf{X}}_E'\mathbf{B}_{iI}^N\right). \end{aligned} \quad \text{(C.1)}$$

C.6 Properties of RWCAL estimator of domain total

We derive the bias under the design and the included/excluded allocation mechanism of the theoretical RWCAL estimator given in (4.33). First, we define the vector of allocation variables for domain i, $\mathbf{c}_i = (c_{i1}, \ldots, c_{iN_i})'$, such that

$$c_{ij} = \begin{cases} 1 & \text{if } j \in U_{iI}; \quad \text{with probability } p_{ij}, \\ 0 & \text{if } j \in U_{iE}; \quad \text{with probability } 1 - p_{ij}. \end{cases}$$

The bias under the design and the allocation distribution is then given by

$$B_{\pi,p}(\tilde{Y}_i^{RWCAL}) = E_p E_{\pi|p}(\tilde{Y}_i^{RWCAL}) - Y_i$$

The expectation under the design and the allocation distribution is

$$\begin{aligned}
& E_p E_{\pi|p}\left[\sum_{j\in s_i}\frac{y_{ij}}{\pi_{ij}p_{ij}} + \left(\sum_{j=1}^{N_i}\mathbf{x}_{ij} - \sum_{j\in s_i}\frac{\mathbf{x}_{ij}}{\pi_{ij}p_{ij}}\right)'\mathbf{B}_i\right] \\
= \; & E_p E_{\pi|p}\left[\sum_{j\in U_{iI}}\frac{y_{ij}}{\pi_{ij}p_{ij}}I_{ij}(s) + \left(\sum_{j=1}^{N_i}\mathbf{x}_{ij} - \sum_{j\in U_{iI}}\frac{\mathbf{x}_{ij}}{\pi_{ij}p_{ij}}I_{ij}(s)\right)'\mathbf{B}_i\right] \\
= \; & E_p\left[\sum_{j\in U_{iI}}\frac{y_{ij}}{\pi_{ij}p_{ij}}E_{\pi|p}(I_{ij}(s)) + \left(\sum_{j=1}^{N_i}\mathbf{x}_{ij} - \sum_{j\in U_{iI}}\frac{\mathbf{x}_{ij}}{\pi_{ij}p_{ij}}E_{\pi|p}(I_{ij}(s))\right)'\mathbf{B}_i\right]. \\
= \; & E_p\left[\sum_{j\in U_{iI}}\frac{y_{ij}}{p_{ij}} + \left(\sum_{j=1}^{N_i}\mathbf{x}_{ij} - \sum_{j\in U_{iI}}\frac{\mathbf{x}_{ij}}{p_{ij}}\right)'\mathbf{B}_i\right] \\
= \; & E_p\left[\sum_{j=1}^{N_i}\frac{y_{ij}}{p_{ij}}c_{ij} + \left(\sum_{j=1}^{N_i}\mathbf{x}_{ij} - \sum_{j=1}^{N_i}\frac{\mathbf{x}_{ij}}{p_{ij}}c_{ij}\right)'\mathbf{B}_i\right] \\
= \; & \sum_{j=1}^{N_i}\frac{y_{ij}}{p_{ij}}E_p(c_{ij}) + \left(\sum_{j=1}^{N_i}\mathbf{x}_{ij} - \sum_{j=1}^{N_i}\frac{\mathbf{x}_{ij}}{p_{ij}}E_p(c_{ij})\right)'\mathbf{B}_i.
\end{aligned}$$

Since, $E_p(c_{ij}) = p_{ij}$ then,

$$\begin{aligned}
E_p E_{\pi|p}(\tilde{Y}_i^{RWCAL}) &= \sum_{j=1}^{N_i} y_{ij} + \left(\sum_{j=1}^{N_i}\mathbf{x}_{ij} - \sum_{j=1}^{N_i}\mathbf{x}_{ij}\right)'\mathbf{B}_i \\
&= \sum_{j=1}^{N_i} y_{ij} = Y_i.
\end{aligned}$$

C.7 Derivation of generalized calibration weights

We obtain the generalized calibration weights $\tilde{h}_{ij}$, which have the form $\tilde{h}_{ij} = w_{ij}(1 + \boldsymbol{\lambda}_i'\mathbf{z}_{ij})$. Replacing these weights in the calibration equation at domain level, that is $\sum_{j\in s_i} \tilde{h}_{ij}\mathbf{x}_{ij} = \mathbf{X}_i$, we obtain

$$\begin{aligned}
\sum_{j\in s_i} \tilde{h}_{ij}\mathbf{x}_{ij} &= \mathbf{X}_i \\
\Leftrightarrow \sum_{j\in s_i} w_{ij}(1 + \boldsymbol{\lambda}_i'\mathbf{z}_{ij})\mathbf{x}_{ij} &= \mathbf{X}_i \\
\Leftrightarrow \sum_{j\in s_i} w_{ij}\mathbf{x}_{ij} + \sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\boldsymbol{\lambda}_i'\mathbf{z}_{ij} &= \mathbf{X}_i \\
\Leftrightarrow \sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{z}_{ij}'\boldsymbol{\lambda}_i &= \mathbf{X}_i - \hat{\mathbf{X}}_i \\
\Leftrightarrow \left(\sum_{j\in s_i} w_{ij}\mathbf{x}_{ij}\mathbf{z}_{ij}'\right)^{-1} (\mathbf{X}_i - \hat{\mathbf{X}}_i) &= \boldsymbol{\lambda}_i.
\end{aligned}$$

Finally, replacing $\boldsymbol{\lambda}_i$ in $\tilde{h}_{ij} = w_{ij}(1 + \boldsymbol{\lambda}_i'\mathbf{z}_{ij})$, we obtain the generalized calibration weights, $\tilde{h}_{ij}$, which are given by

$$\tilde{h}_{ij} = w_{ij}\left[1 + (\mathbf{X}_i - \hat{\mathbf{X}}_i)'\left(\sum_{j\in s_i} w_{ij}\mathbf{z}_{ij}\mathbf{x}_{ij}'\right)^{-1}\mathbf{z}_{ij}\right].$$

C.8 Design-bias of BLUP of domain mean under cut-off sampling with srswor within the set of included units

Consider the theoretical BLUP for negligible sampling fraction, given by

$$\tilde{\bar{Y}}_i^{BLUP} = \gamma_{is}[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\boldsymbol{\beta}] + (1 - \gamma_{is})\bar{\mathbf{X}}_i'\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the true regression parameter in model (4.38), considering that this model holds for all the units of the population, either included or excluded. The design-bias

of this theoretical BLUP is given by

$$\begin{aligned}
B_\pi(\tilde{\bar{Y}}_i^{BLUP}) &= E_\pi(\tilde{\bar{Y}}_i^{BLUP}) - \bar{Y}_i \\
&= E_\pi\{\gamma_{is}[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\boldsymbol{\beta}] + (1+\gamma_{is})\bar{\mathbf{X}}_i'\boldsymbol{\beta}\} - \gamma_{is}\bar{Y}_i - (1-\gamma_{is})\bar{Y}_i \\
&= \gamma_{is}[\bar{Y}_{iI} + (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_{iI})'\boldsymbol{\beta}] + (1-\gamma_{is})\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \gamma_{is}\bar{Y}_i - (1-\gamma_{is})\bar{Y}_i \\
&= \gamma_{is}[\bar{Y}_{iI} + (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_{iI})'\boldsymbol{\beta} - \bar{Y}_i] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i) \\
&= \gamma_{is}[(\bar{Y}_{iI} - \bar{\mathbf{X}}_{iI}'\boldsymbol{\beta}) + (\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i)] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i) \\
&= \gamma_{is}\left[\frac{1}{N_{iI}}(Y_{iI} - \mathbf{X}_{iI}'\boldsymbol{\beta}) + \frac{1}{N_i}(\mathbf{X}_i'\boldsymbol{\beta} - Y_i)\right] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i).
\end{aligned}$$

Using now $Y_{iI} = Y_i - Y_{iE}$ and similarly for the total $\mathbf{X}_{iI}$, we obtain

$$\begin{aligned}
B_\pi(\tilde{\bar{Y}}_i^{BLUP}) &= \gamma_{is}\left[\frac{1}{N_{iI}}(Y_i - \mathbf{X}_i'\boldsymbol{\beta} - Y_{iE} + \mathbf{X}_{iE}'\boldsymbol{\beta}) + \frac{1}{N_i}(\mathbf{X}_i'\boldsymbol{\beta}_I - Y_i)\right] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i) \\
&= \gamma_{is}\left[\left(\frac{1}{N_i} - \frac{1}{N_{iI}}\right)(\mathbf{X}_i'\boldsymbol{\beta} - Y_i) - \frac{1}{N_{iI}}(Y_{iE} - \mathbf{X}_{iE}'\boldsymbol{\beta})\right] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i) \\
&= \gamma_{is}\left[\left(\frac{N_{iE}}{N_i N_{iI}}\right)(Y_i - \mathbf{X}_i'\boldsymbol{\beta}) - \frac{1}{N_{iI}}(Y_{iE} - \mathbf{X}_{iE}'\boldsymbol{\beta})\right] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i) \\
&= \gamma_{is}\frac{N_{iE}}{N_{iI}}\left[(\bar{Y}_i - \bar{\mathbf{X}}_i'\boldsymbol{\beta}) - (\bar{Y}_{iE} - \bar{\mathbf{X}}_{iE}'\boldsymbol{\beta})\right] + (1-\gamma_{is})(\bar{\mathbf{X}}_i'\boldsymbol{\beta} - \bar{Y}_i).
\end{aligned}$$

The design-variance of $\tilde{\bar{Y}}_i$ is given by

$$\begin{aligned}
V_\pi(\hat{\bar{Y}}_i^{BLUP}) &= V_\pi\{\gamma_i[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\boldsymbol{\beta}] + (1+\gamma_i)\bar{\mathbf{X}}_i'\boldsymbol{\beta}\} \\
&= V_\pi[\gamma_i(\bar{y}_{is} - \hat{\bar{\mathbf{x}}}_{is}'\boldsymbol{\beta})] \\
&= V_\pi(\gamma_i\bar{\varepsilon}_i),
\end{aligned}$$

for $\bar{\varepsilon}_i = \bar{y}_{is} - \bar{\mathbf{x}}_{is}'\boldsymbol{\beta}$. Then, using the basic design-based inference, we obtain

$$\begin{aligned}
V_\pi(\gamma_i\bar{\varepsilon}_i) &= V_\pi\left(\frac{\gamma_i}{N_i}\sum_{j\in s_i} w_{ij}\varepsilon_{ij}\right) \\
&= \frac{\gamma_i^2}{N_i^2}V_\pi\left(\sum_{j\in s_i} w_{ij}\varepsilon_{ij}\right) \cong \gamma_i^2 V_\pi(\tilde{\bar{Y}}_i^{LCAL}).
\end{aligned}$$

Note that, if model (4.38) holds for the included and excluded units, then $E_m(\mathbf{B}_{iI}) = \boldsymbol{\beta}$ and $V_m(\mathbf{B}_{iI}) = (\sum_{j\in U_i}\mathbf{x}_{ij}\mathbf{x}_{ij}')^{-1}$, which is $O(N_i^{-1})$ under standard regularity conditions. Since typically the population size N_i is large, we have $\mathbf{B}_{iI} \cong \boldsymbol{\beta}$ under the model.

C.9 Cut-off sampling versus srswor

Here we show that the estimators obtained by taking the whole set of included units as the sample (with no sampling after cut-off) take the same form as those obtained considering that the set of included units is a srswor from the whole population. However, of course, the properties of the estimators are not the same under these two sampling schemes, as we have shown in this work.

First, under cut-off sampling, taking the whole set of included units as the sample ($s_i = U_{iI}$), the area sample size n_i is equal to the area population size of U_{iI}, that is, $n_i = N_{iI}$. In this case, $w_{ij} = 1$ for all $j \in U_{iI}$. The direct (HT) estimator for the area total Y_i under this set up reduces to

$$\hat{Y}_i = \sum_{j \in s_i} y_{ij} = \sum_{j \in U_{iI}} y_{ij} = Y_{iI}. \tag{C.2}$$

Similarly, the LCAL estimator reduces to

$$\begin{aligned}
\hat{Y}_i^{LCAL} &= \hat{Y}_i^{GREG} = \hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)'\hat{\mathbf{B}}_i, \\
&= \sum_{j \in s_i} w_{ij} y_{ij} + \left(\mathbf{X}_i - \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \right)' \left(\sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} y_{ij} \\
&= \sum_{j \in U_{iI}} y_{ij} + \left(\mathbf{X}_i - \sum_{j \in U_{iI}} \mathbf{x}_{ij} \right)' \left(\sum_{j \in U_{iI}} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \sum_{j \in U_{iI}} \mathbf{x}_{ij} y_{ij} \\
&= Y_{iI} + (\mathbf{X}_i - \mathbf{X}_{iI})'\mathbf{B}_{iI},
\end{aligned} \tag{C.3}$$

where $\mathbf{X}_{iI} = \sum_{j \in U_{iI}} \mathbf{x}_{ij}$ and $\mathbf{B}_{iI} = \left(\sum_{j \in U_{iI}} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \sum_{j \in U_{iI}} \mathbf{x}_{ij} y_{ij}$.

Finally, the BLUP of Y_i under this setup reduces to

$$\begin{aligned}
\hat{Y}_i^{BLUP} &= \gamma_{is} N_i \left[\frac{\sum_{j \in s_i} y_{ij}}{n_i} + \left(\bar{\mathbf{X}}_i - \frac{\sum_{j \in s_i} \mathbf{x}_{ij}}{n_i} \right)' \tilde{\boldsymbol{\beta}}_s \right] + (1 - \gamma_{is}) N_i \bar{\mathbf{X}}_i' \tilde{\boldsymbol{\beta}}_s \\
&= \gamma_{is} N_i \left[\frac{\sum_{j \in U_{iI}} y_{ij}}{N_{iI}} + \left(\bar{\mathbf{X}}_i - \frac{\sum_{j \in U_{iI}} \mathbf{x}_{ij}}{N_{iI}} \right)' \tilde{\boldsymbol{\beta}}_s \right] + (1 - \gamma_{is}) N_i \bar{\mathbf{X}}_i' \tilde{\boldsymbol{\beta}}_s \\
&= \gamma_{is} \left[N_i \bar{Y}_{iI} + (\mathbf{X}_i - N_i \bar{\mathbf{X}}_{iI})' \tilde{\boldsymbol{\beta}}_s \right] + (1 - \gamma_{is}) \mathbf{X}_i' \tilde{\boldsymbol{\beta}}_s.
\end{aligned} \tag{C.4}$$

Consider now that we draw a sample, s_i, from the population U_i using srswor and the sample obtained is U_{iI}, that is, $s_i = U_{iI}$. The sample size is N_{iI} and the sampling weights are then $w_{ij} = N_i/N_{iI}$ for all $j \in U_{iI}$. The direct estimator of the domain total

Y_i becomes

$$\hat{Y}_i = \sum_{j \in s_i} y_{ij} = \sum_{j \in U_{iI}} y_{ij} = Y_{iI},$$

which is equal to the direct estimator in (C.2). The direct LCAL estimator becomes

$$\begin{aligned}
\hat{Y}_i^{LCAL} &= \hat{Y}_i^{GREG} = \hat{Y}_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)'\hat{\mathbf{B}}_i, \\
&= \sum_{j \in s_i} w_{ij} y_{ij} + \left(\mathbf{X}_i - \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \right)' \left(\sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} y_{ij} \\
&= \sum_{j \in s_{iI}} N_i/N_{iI} y_{ij} + \left(\mathbf{X}_i - \sum_{j \in s_{iI}} N_i/N_{iI} \mathbf{x}_{ij} \right)' \left(\sum_{j \in s_{iI}} N_i/N_{iI} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \sum_{j \in s_{iI}} N_i/N_{iI} \mathbf{x}_{ij} y_{ij} \\
&= \sum_{j \in U_{iI}} y_{ij} + \left(\mathbf{X}_i - \sum_{j \in U_{iI}} \mathbf{x}_{ij} \right)' \left(\sum_{j \in U_{iI}} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \sum_{j \in U_{iI}} \mathbf{x}_{ij} y_{ij} \\
&= Y_{iI} + (\mathbf{X}_i - \mathbf{X}_{iI})'\mathbf{B}_{iI}
\end{aligned}$$

which is the same as (C.3). The same occurs for calibration after reweighting and generalized calibration. Finally, the BLUP estimator becomes

$$\begin{aligned}
\hat{Y}_i^{BLUP} &= \gamma_{is} N_i \left[\frac{\sum_{j \in s_i} y_{ij}}{n_i} + \left(\bar{\mathbf{X}}_i - \frac{\sum_{j \in s_i} \mathbf{x}_{ij}}{n_i} \right)' \tilde{\boldsymbol{\beta}}_s \right] + (1 - \gamma_{is}) N_i \bar{\mathbf{X}}_i' \tilde{\boldsymbol{\beta}}_s \\
&= \gamma_{is} N_i \left[\frac{\sum_{j \in s_i} N_i/N_{iI} y_{ij}}{n_i} + \left(\bar{\mathbf{X}}_i - \frac{\sum_{j \in s_i} N_i/N_{iI} \mathbf{x}_{ij}}{n_i} \right)' \tilde{\boldsymbol{\beta}}_s \right] + (1 - \gamma_{is}) N_i \bar{\mathbf{X}}_i' \tilde{\boldsymbol{\beta}}_s \\
&= \gamma_{is} N_i \left[\frac{\sum_{j \in U_{iI}} y_{ij}}{N_{iI}} + \left(\bar{\mathbf{X}}_i - \frac{\sum_{j \in U_{iI}} \mathbf{x}_{ij}}{N_{iI}} \right)' \tilde{\boldsymbol{\beta}}_s \right] + (1 - \gamma_{is}) N_i \bar{\mathbf{X}}_i' \tilde{\boldsymbol{\beta}}_s \\
&= \gamma_{is} \left[N_i \bar{Y}_{iI} + (\mathbf{X}_i - N_i \bar{\mathbf{X}}_{iI})' \tilde{\boldsymbol{\beta}}_s \right] + (1 - \gamma_{is}) \mathbf{X}_i' \tilde{\boldsymbol{\beta}}_s,
\end{aligned}$$

which is the same as in (C.4).

C.10 Estimates of total sales by provinces

Table C.1: Direct, GREG and EB estimates of total sales for the selected product and estimated coefficients of variation (%) for each Spanish province. Results sorted by increasing sample size.

PROVINCE	n_i	$\hat{Y}_i^{DIR}$	$\hat{Y}_i^{GREG}$	$\hat{Y}_i^{EB}$	$cv(\hat{Y}_i^{DIR})$	$cv(\hat{Y}_i^{GREG})$	$cv(\hat{Y}_i^{EB})$
SORIA	3	293020.0	187824.9	213325.0	50.0	17.1	6.2
ZAMORA	7	932520.0	345095.8	454657.0	43.3	18.9	5.5
ALAVA	11	130083.6	119918.5	118835.3	23.7	14.7	9.7
ALMERIA	13	1870104.6	2407333.1	2272051.4	30.4	5.8	3.4
PALENCIA	14	626340.0	380367.4	409775.4	16.7	7.6	4.1
SALAMANCA	14	1265580.0	966094.1	1068230.6	21.9	7.3	3.9
AVILA	15	708696.0	392474.1	418917.2	19.5	9.2	5.0
LERIDA	17	817817.6	1011032.3	1014770.2	22.5	7.1	4.1
CIUDAD REAL	18	1764000.0	841228.2	939994.9	21.4	8.6	4.6
GUADALAJARA	18	463047.8	362148.3	363856.9	17.1	6.0	4.5
RIOJA	18	809900.0	622488.3	595178.6	18.2	5.2	3.7
SEGOVIA	19	610370.5	386734.4	402324.0	15.7	7.5	4.2
CACERES	20	4391826.0	2081619.7	2286462.0	20.4	5.6	2.7
GUIPUZCOA	20	181634.0	136700.0	156311.8	18.6	16.7	11.6
HUESCA	22	377954.5	372101.3	371246.5	24.5	7.7	5.2
TERUEL	22	534417.3	446565.7	465643.3	19.9	6.0	4.3
CUENCA	23	588464.3	587005.5	586347.5	19.0	5.8	4.2
VALLADOLID	24	1609875.0	1210132.8	1188336.1	13.3	4.5	3.4
BURGOS	28	961645.7	708510.0	666698.1	18.5	4.9	3.4
CORDOBA	28	4457614.3	3367169.5	3312801.5	17.9	3.4	2.4
ORENSE	28	148577.1	88104.6	108428.9	17.4	19.0	10.5
LUGO	30	107213.3	92938.7	104233.7	16.9	13.8	10.7
ALBACETE	31	1654606.5	1115182.2	1073719.8	13.4	4.2	2.8
LEON	31	1528254.2	1274531.6	1270341.6	14.5	4.2	3.2
HUELVA	32	3031328.1	2838874.0	2816281.3	10.5	2.6	2.0
NAVARRA	33	1291343.0	956737.9	957660.4	13.2	4.4	3.4
PONTEVEDRA	33	159229.1	107198.9	138367.4	22.2	19.7	13.4
VIZCAYA	34	228618.8	183267.3	206304.6	13.1	13.2	9.1
TOLEDO	35	1619939.4	1529104.8	1539799.3	13.1	4.2	3.2
CADIZ	38	1851521.1	1585755.9	1620844.2	14.9	4.0	3.4
BADAJOZ	39	4571743.6	3439625.5	3457692.5	13.5	2.7	2.2

Continued on next page

Table C.1 – *Continued from previous page*

PROVINCE	n_i	$\hat{Y}_i^{DIR}$	$\hat{Y}_i^{GREG}$	$\hat{Y}_i^{EB}$	$cv(\hat{Y}_i^{DIR})$	$cv(\hat{Y}_i^{GREG})$	$cv(\hat{Y}_i^{EB})$
MALAGA	39	2499392.3	3188031.1	3237081.8	10.9	4.2	2.5
TARRAGONA	41	2872882.0	2690969.7	2656117.8	11.6	2.6	2.2
GRANADA	42	2123693.3	2221155.1	2241916.2	12.5	3.8	2.9
JAEN	43	1928229.8	1940379.2	1943101.0	15.8	3.2	2.7
ZARAGOZA	43	3750210.7	2564909.0	2578011.3	13.5	3.0	2.3
GERONA	45	2029222.2	1748165.7	1767490.3	10.4	3.2	2.5
MURCIA	51	6700070.6	7467465.0	7341434.6	8.7	2.2	1.6
BALEARES	52	849950.8	650012.6	694416.3	21.5	6.1	4.7
CANTABRIA	52	285632.3	204947.7	226163.1	10.7	9.5	6.4
ASTURIAS	55	2113034.5	1702020.8	1661932.8	13.5	3.6	3.1
CASTELLON	55	1605604.4	1526618.1	1530394.2	8.9	2.5	2.2
SEVILLA	55	7458078.2	6878368.2	6857368.8	11.0	2.0	1.7
CORUNA	62	340200.0	217028.5	206041.8	20.2	10.9	10.2
ALICANTE	66	8324589.1	8390895.3	8240996.9	9.2	1.8	1.6
VALENCIA	113	7671137.7	7209128.2	7153290.2	6.3	1.7	1.4
MADRID	123	11483342.8	12892853.8	12892305.0	6.2	1.7	1.5
BARCELONA	187	22356500.5	24990558.9	24797372.9	4.8	1.0	0.9